uOttawa

A Deployment-Oriented Evaluation of Hybrid Deep Learning and Traditional Signal Processing for Acoustic Echo Cancellation and Speech Enhancement

Aidan Charles

Supervisor
Dr. Martin Bouchard

Thesis submitted to the University of Ottawa in partial fulfillment of the requirements for the degree of
Master of Applied Science in
Electrical and Computer Engineering
with Concentration in Applied
Artificial Intelligence

School of Electrical Engineering and Computer Science
Faculty of Engineering, University of Ottawa

© Aidan Charles, Ottawa, Canada, 2026

Abstract

Acoustic Echo Cancellation (AEC) is a key component of hands-free communication systems, enabling full-duplex operation by suppressing acoustic feedback between loudspeakers and microphones. Traditional adaptive filtering-based AEC systems remain widely used due to their efficiency and robustness but are limited by residual echo, nonlinear distortions, and background noise. Although hybrid deep learning-based enhancement methods have shown strong performance, they are often evaluated using idealized front-end configurations, making it difficult to isolate their contribution in real-time systems. This thesis presents a hybrid AEC framework that combines a production-oriented frequency-domain adaptive filter with a deep learning post-filter. The proposed model is derived from a convolutional recurrent architecture and redesigned for deployment using a Complex Ratio Mask (CRM)-based enhancement strategy, a simplified decoder, and Frequency-Temporal Long Short-Term Memory (F-T-LSTM) layers for temporal modelling. An identical linear echo cancellation front-end is used for both a traditional cascaded residual echo and noise suppression pipeline and the proposed neural system, enabling a controlled comparison. Evaluation is conducted using standardized ICASSP benchmark scenarios, including near-end single-talk, far-end single-talk, and double-talk conditions. Performance is assessed using perceptual metrics for echo suppression, speech distortion, and noise reduction, while Echo Return Loss Enhancement is used to evaluate suppression dynamics. Results show that the proposed system achieves comparable echo suppression to the traditional pipeline while improving speech preservation and overall perceptual quality. It also adapts more rapidly to changing acoustic

conditions, whereas the baseline exhibits slower recovery. The proposed model achieved the highest perceptual evaluation scores, including an Echo Mean Opinion Score (MOS) of 4.53, a Degradation MOS of 3.55, and an overall perceptual MOS of 3.90. The system operates with 30 ms latency and a real-time factor of 0.68, demonstrating near real-time performance suitable for deployment.

Acknowledgments

I would like to express my sincere gratitude to my supervisor, Dr. Bouchard, for his guidance, mentorship, and encouragement throughout this journey. His insight helped shape the direction of this thesis, and his support, including fostering valuable industry collaborations, was instrumental to its success. I am grateful for the opportunity to learn from him throughout this challenging and rewarding process.

I would also like to thank Hala, Michael, and Sa'di for their contributions to this work. Their technical expertise, guidance, and support during the development of the final system were invaluable, and their feedback helped strengthen both the research and the overall direction of this thesis.

I would like to dedicate a special acknowledgement to the memory of Richard. His efforts in facilitating the collaboration that made this work possible had a lasting impact, and this thesis would not have been possible without his support and commitment.

Finally, I would like to thank my family for their unwavering encouragement and support throughout my academic journey. Their belief in me gave me the confidence to take the next step in pursuing a master's degree and helped me navigate the many challenges and triumphs along the way.

To all those who contributed to this work, thank you.

Contents

Abstract ii

Acknowledgments iv

Contents v

List of Tables xi

List of Figures xii

List of Acronyms xv

1 Introduction 1

1.1 Introduction 1

1.2 Classical AEC Systems 2

1.3 Deep Learning for Speech Enhancement and Echo Cancellation 2

1.4 Hybrid Signal Processing and Deep Learning Approaches 3

1.5 Motivation and Problem Statement 4

1.6 Contributions of This Thesis 6

1.7 Thesis Organization 8

2 Background 9

2.1 Acoustic Echo Path 10

2.1.1 Signal Model 10
2.1.2 Room Impulse Response (RIR) 12
2.1.3 Time-Varying Nature of the Echo Path 12
2.1.4 Challenges in Echo Modeling 13
2.2 Time-Frequency Processing and Block Processing 14
2.2.1 Short-Time Fourier Transform (STFT) 14
2.2.2 Magnitude and Phase Representation 15
2.2.3 Block Processing for Real-Time Systems 16
2.3 Linear AEC Algorithms 17
2.3.1 Overview of Linear AEC 17
2.4 Residual Echo Suppression (RES) 21
2.4.1 Overview 21
2.4.2 Spectral Domain Formulation 22
2.4.3 Wiener Filtering Approach 23
2.4.4 Spectral Gain-Based Methods 24
2.5 Noise Suppression (Noise Suppression (NS)) 27
2.5.1 Overview 27
2.5.2 Classical Noise Suppression Methods 28
2.5.3 Interaction with Residual Echo Suppression 29
2.6 Deep Learning for Speech Enhancement 30
2.6.1 Overview 30
2.6.2 Convolutional Neural Network (CNN) Based Processing 32
2.6.3 Encoder-Decoder Architectures 33
2.6.4 Sequence Modeling with Recurrent Networks 35
2.6.5 Mask-Based and Spectral Mapping Approaches 37

3 Related Works **39**

3.1 Classical Formulations of AEC, Acoustic Echo Suppression (AES), and NS . 40

3.1.1 Classical AEC 40
3.1.2 Classical NS and AES 41
3.1.3 Limitations of Classical Methods 42
3.2 Deep Learning for Speech Enhancement and Echo Suppression 43
3.2.1 Spectral Masking Approaches 45
3.2.2 Complex-Domain Speech Enhancement Methods 46
3.2.3 Architectural Evolution of Deep Learning-Based Enhancement Systems 48
3.3 Research Gap and Proposed Contribution 51

4 Proposed Hybrid AEC System 55
4.1 System Overview 55
4.2 Linear AEC Stage 57
4.2.1 Partitioned Block Frequency-Domain Adaptive Filter (PBFDAF) Linear Acoustic Echo Cancellation 57
4.2.2 Partitioning of the Adaptive Filter 59
4.2.3 Block-Based Input Representation 61
4.2.4 Frequency-Domain Filtering 63
4.2.5 Error Signal Computation and Adaptive Filter Update 64
4.2.6 Practical Justification Compared to Multidelay Block Frequency Domain (MDF) and Recursive Least Squares (RLS) 65
4.2.7 Summary of the PBFDAF Processing Chain 67
4.3 Deep Learning Post-Filter 69
4.3.1 Feature Extraction and Signal Representation 70
4.3.2 Architecture 72
4.3.3 Output Formulation 90
4.4 Loss Functions 91
4.4.1 Signal Reconstruction for Loss Computation 92
4.4.2 Voice Activity-Aware Weighting 92

4.4.3 Magnitude Loss with Asymmetric Penalty 93
4.4.4 Power-Law Compressed Phase-Aware Loss 95
4.4.5 Scale-Invariant Signal-to-Distortion Ratio (SI-SDR) Loss 95
4.4.6 Complex Mask Regularization (Bias Term) 96
4.4.7 Overall Loss Function 97
4.5 Baseline System 98
4.5.1 Traditional Processing Pipeline 98
4.5.2 Wiener-Based Suppression 99
4.5.3 Limitations of the Baseline 103

5 Experimental Setup 105
5.1 Training Dataset 105
5.2 Training Preprocessing 108
5.3 Training Procedure 110
5.3.1 Optimization Setup 110
5.3.2 Loss Configuration 111
5.3.3 Batching and Data Splitting 111
5.3.4 Chunk-wise Training Strategy 112
5.3.5 Training Loop and Convergence 113
5.3.6 Validation Strategy 113
5.3.7 Implementation Details 114
5.4 Test Dataset 114
5.5 Testing Procedure 116
5.6 Evaluation Metrics 117
5.6.1 Acoustic Echo Cancellation Mean Opinion Score (AECMOS) 118
5.6.2 Deep Noise Suppression Mean Opinion Score (DNSMOS) 119
5.6.3 Echo Return Loss Enhancement (ERLE) 120

6 Results and Analysis **123**

6.1 Introduction . . . 123

6.2 Objective Results . . . 124

6.2.1 AECMOS Results . . . 124

6.2.2 DNSMOS Results . . . 126

6.2.3 Overall Trends . . . 128

6.3 Scenario-Based Performance Analysis . . . 129

6.3.1 Traditional Signal Processing Pipeline Analysis . . . 130

6.4 Temporal Behaviour Analysis (Single-File ERLE) . . . 141

6.4.1 Linear AEC Baseline . . . 141

6.4.2 Linear AEC + AES . . . 143

6.4.3 Linear AEC + AES + NS . . . 144

6.4.4 Proposed Hybrid Deep Learning Model . . . 146

6.4.5 Overall Temporal Trends . . . 148

6.5 Computational Performance and Real-Time Considerations . . . 149

6.5.1 Latency Analysis . . . 149

6.5.2 Real Time Factor (RTF) Evaluation . . . 151

6.5.3 Discussion . . . 152

6.6 Ablation Study . . . 153

6.6.1 Architectural Variations . . . 154

6.6.2 Mask Representation . . . 159

6.6.3 Loss Function Design . . . 160

6.6.4 Input Feature Configuration . . . 165

6.6.5 Temporal Context and Sliding Window Configuration . . . 166

6.6.6 Stateful vs Non-Stateful Processing . . . 168

6.6.7 Training Hyperparameters . . . 169

6.7 Discussion . . . 170

6.7.1 Performance Trade-offs in Echo Cancellation 170
6.7.2 Temporal Modelling and Dynamic Behaviour 171
6.7.3 Advantages of the Hybrid Learning Approach 172
6.7.4 Real-Time Considerations and Practical Deployment 173
6.7.5 Overall System Insights 174

7 Conclusion and Future Work **175**
7.1 Conclusion 175
7.2 Future Work 178

References **181**

List of Tables

4.1 Encoder architecture showing progressive channel expansion and time–frequency downsampling. 78

4.2 Decoder architecture consisting of transposed convolutional upsampling stages followed by a final 1×1 output projection layer producing the complex ratio mask representation. 88

5.1 Evaluation scenarios used from the International Conference on Acoustics, Speech, and Signal Processing (ICASSP) AEC Challenge test dataset. 115

6.1 AECMOS results across different acoustic scenarios 125

6.2 DNSMOS results across different acoustic scenarios 126

List of Figures

2.1 Acoustic echo signal model illustrating the formation of the microphone signal from the far-end signal, acoustic echo path, near-end speech, and background noise. 11

2.2 Block diagram of a linear AEC system. The adaptive filter estimates the echo signal using the far-end reference signal, and the resulting residual error signal is used to adapt the filter coefficients. 19

2.3 Generic deep learning-based speech enhancement pipeline operating in the time-frequency domain. A corrupted speech signal is transformed into a spectral representation using the STFT, processed by an encoder–bottleneck–decoder architecture incorporating temporal modeling Recurrent Neural Network (RNN), and mapped to an enhanced speech spectrogram. 31

4.1 Overall hybrid AEC system architecture. The pipeline consists of a Generalized Cross-Correlation with Phase Transform (GCC-PHAT) alignment stage, a PBFDAF, and a deep learning post-filter operating in the time-frequency domain. The neural model consists of a convolutional encoder, F-T-LSTM bottleneck, and decoder producing a CRM. 56

4.2 Structure of the Gated Convolutional (GConv) block used throughout the encoder. The feature branch extracts spectral representations while the gating branch learns a signal-dependent mask that selectively controls information flow. 74

4.3 Structure of the proposed F-T-LSTM layer. The module first applies bidirectional recurrent modeling along the frequency axis, followed by temporal modeling along the time axis, enabling joint spectral–temporal representation learning for speech enhancement and AEC. 82

4.4 Structure of the Transposed Gated Convolutional (TransGConv) block used in the decoder. Transposed convolutions perform learned upsampling while a gating branch adaptively regulates feature flow during reconstruction. 87

5.1 Overview of the synthetic data generation process used in the ICASSP AEC Challenge dataset. Far-end speech is convolved with a synthetic RIR to generate acoustic echo, while near-end speech and optional background noise are combined to create realistic microphone mixtures. The clean near-end speech is retained as the ground-truth target for supervised training. 107

6.1 Overall performance comparison across all evaluated systems. The proposed hybrid model achieves the best balance between echo suppression, speech preservation, and perceptual quality, outperforming both the traditional cascaded pipeline and intermediate baselines. 129

6.2 AECMOS and DNSMOS results for the baseline system across near-end single-talk, far-end single-talk, double-talk, and overall conditions. 130

6.3 AECMOS and DNSMOS results for the linear AEC system across near-end single-talk, far-end single-talk, double-talk, and overall conditions. 132

6.4 AECMOS and DNSMOS results for the AEC + AES system across near-end single-talk, far-end single-talk, double-talk, and overall conditions. 135

6.5 AECMOS and DNSMOS results for the AEC + AES + NS system across near-end single-talk, far-end single-talk, double-talk, and overall conditions. . 137

6.6 AECMOS and DNSMOS results for the proposed hybrid deep learning model across near-end single-talk, far-end single-talk, double-talk, and overall conditions. 139

6.7 ERLE over time for linear AEC using proprietary PBFDAF Algorithm . . . 142

6.8 ERLE over time for linear AEC with AES 143

6.9 ERLE over time for linear AEC + AES + NS 145

6.10 ERLE over time for proposed hybrid deep learning model 147

List of Acronyms

ADSP Acoustic Digital Signal Processing

AEC Acoustic Echo Cancellation

AECMOS Acoustic Echo Cancellation Mean Opinion Score

AES Acoustic Echo Suppression

BAK Background

BN Batch Normalization

CNN Convolutional Neural Network

COLA Constant Overlap-Add

CPU Central Processing Unit

CRM Complex Ratio Mask

CRN Convolutional Recurrent Network

DC Direct Current

DCCRN Deep Complex Convolution Recurrent Network

DEMAND Diverse Environments Multi-channel Acoustic Noise Database

DNSMOS Deep Noise Suppression Mean Opinion Score

DTD Double-Talk Detection

DRC Dynamic Range Compression

DSP Digital Signal Processing

ERLE Echo Return Loss Enhancement

FIR Finite Impulse Response

FFT Fast Fourier Transform

F-LSTM Frequency Long Short-Term Memory

F-T-LSTM Frequency-Temporal Long Short-Term Memory

GConv Gated Convolutional

GCC-PHAT Generalized Cross-Correlation with Phase Transform

GPU Graphics Processing Unit

IBM Ideal Binary Mask

ICASSP International Conference on Acoustics, Speech, and Signal Processing

IRM Ideal Ratio Mask

IFFT Inverse Fast Fourier Transform

ISTFT Inverse Short-Time Fourier Transform

LMS Least Mean Squares

LSTM Long Short-Term Memory

MDF Multidelay Block Frequency Domain

MMSE Minimum Mean Squared Error

MSE Mean Squared Error

MOS Mean Opinion Score

NLMS Normalized Least Mean Squares

NS Noise Suppression

OLA Overlap-Add

OLS Overlap-Save

OVRL Overall

PBFDAF Partitioned Block Frequency-Domain Adaptive Filter

PSD Power Spectral Density

PSM Phase-Sensitive Mask

PReLU Parametric Rectified Linear Unit

ReLU Rectified Linear Unit

RES Residual Echo Suppression

RIR Room Impulse Response

RLS Recursive Least Squares

RNN Recurrent Neural Network

RTF Real Time Factor

RT60 Reverberation Time

SI-SDR Scale-Invariant Signal-to-Distortion Ratio

SIG Signal

SNR Signal-to-Noise Ratio

STFT Short-Time Fourier Transform

TransGConv Transposed Gated Convolutional

T-LSTM Temporal Long Short-Term Memory

VAD Voice Activity Detection

wRLS Weighted Recursive Least Squares

Chapter 1

Introduction

1.1 Introduction

Acoustic Echo Cancellation (AEC) is a fundamental component of modern hands-free communication systems, enabling natural and intelligible speech transmission in scenarios where loudspeaker playback is captured by a microphone and reintroduced into the communication loop as undesired echo. This problem is particularly relevant in real-world applications such as automotive communication systems, teleconferencing platforms, smart speakers, and voice-controlled embedded devices, where simultaneous playback and recording are required for full-duplex communication [1].

In these environments, acoustic echo arises due to the coupling between loudspeakers, microphones, and the surrounding acoustic space. Without effective echo cancellation, audio played through the loudspeaker is captured again by the microphone after being reflected and delayed within the environment, significantly degrading communication quality and user experience. As a result, AEC has become a core technology in modern speech processing systems, often deployed alongside noise suppression and speech enhancement modules in real-time audio pipelines.

1.2 Classical AEC Systems

Traditional AEC systems are primarily based on adaptive filtering techniques designed to model and remove the linear component of the echo path. Algorithms such as Least Mean Squares (LMS) and frequency-domain adaptive filters adaptively model the acoustic impulse response between the loudspeaker and microphone, enabling estimation and cancellation of the dominant linear echo component [1, 2].

In practical systems, linear AEC alone is often insufficient. Residual echo and noise components remain due to nonlinearities introduced by loudspeakers, microphone distortions, and complex acoustic reflections. To address these limitations, classical speech enhancement pipelines extend the processing chain with additional modules such as Acoustic Echo Suppression (AES) and Noise Suppression (NS) [3, 4]. These components are typically applied sequentially, forming cascaded architectures where each stage targets a specific degradation source.

While this modular design is computationally efficient and widely deployed in industrial systems, it has inherent limitations [5]. Since each stage is independently optimized, error propagation between modules can occur, and aggressive suppression in later stages may introduce speech distortion or noise artifacts. Furthermore, the separation between echo and noise suppression is often done independently, limiting the ability of the system to jointly optimize perceptual speech quality. To help mitigate these limitation new approaches have been developed using data-driven deep learning based methods.

1.3 Deep Learning for Speech Enhancement and Echo Cancellation

Recent advances in deep learning have significantly transformed speech enhancement and AEC by enabling data-driven approaches capable of learning complex nonlinear relationships

directly from training data [6, 7]. Unlike traditional signal processing methods that rely on handcrafted statistical assumptions, deep neural networks can model residual echo, nonlinear distortions, reverberation, and background noise jointly, leading to substantial improvements in challenging acoustic environments [7, 8].

Neural network-based speech enhancement systems typically operate in either the time domain or the time-frequency domain, leveraging architectures such as Convolutional Neural Networks (CNN), Recurrent Neural Networks (RNN), and encoder-decoder frameworks [9, 10]. Time-frequency approaches commonly employ spectrogram representations combined with convolutional and recurrent layers to estimate spectral masks or enhanced speech spectra, while time-domain methods directly process raw audio waveforms [6]. These approaches have demonstrated strong performance improvements in scenarios involving non-stationary noise, reverberation, and nonlinear echo components [11].

Despite their strong enhancement capability, many deep learning-based approaches are not inherently designed for strict real-time deployment constraints [12]. Large model sizes, computational complexity, memory requirements, refresh rate, and algorithmic latency can limit their integration into embedded or production-grade speech communication systems [13]. Furthermore, many deep learning approaches operate independently from classical adaptive filtering pipelines, reducing compatibility with established industrial AEC frameworks. As a result, recent research has increasingly explored hybrid approaches that combine the robustness of adaptive filtering with the nonlinear modeling capability of deep learning systems.

1.4 Hybrid Signal Processing and Deep Learning Approaches

To address the limitations of both classical and deep learning-based methods, hybrid approaches have emerged that combine adaptive signal processing with neural network-based

enhancement. These systems typically employ a traditional AEC front-end to suppress the dominant linear echo component, followed by neural post-filters or residual enhancement networks designed to suppress remaining echo, noise, and nonlinear distortions [8].

Hybrid systems aim to leverage the complementary strengths of both paradigms: the stability, interpretability, and computational efficiency of adaptive filtering, and the expressive modeling capability of deep learning architectures for complex nonlinear acoustic effects [5]. By combining these approaches, hybrid architectures can better handle real-world conditions where both linear and nonlinear distortions are present simultaneously.

While numerous studies have reported strong performance for hybrid AEC systems that combine adaptive filtering with deep learning-based post-processing [5, 8, 14], most published evaluations compare alternative deep learning architectures, training strategies, or benchmark challenge submissions rather than directly comparing deep learning-based enhancement against mature traditional speech enhancement pipelines under controlled conditions [12]. Consequently, the practical benefits and trade-offs of replacing conventional Residual Echo Suppression (RES) and NS modules with deep learning-based alternatives while maintaining the same linear echo cancellation front-end remain difficult to quantify.

1.5 Motivation and Problem Statement

Despite the promising results reported for hybrid AEC systems, several challenges remain that limit both fair comparison across studies and the assessment of their practical deployment benefits. First, relatively few studies directly compare hybrid deep learning-based enhancement systems against mature traditional speech enhancement pipelines under controlled conditions. Most published evaluations focus on comparisons between alternative neural architectures, training strategies, or benchmark challenge submissions. Therefore, it is often unclear how much practical benefit is obtained by replacing conventional RES and NS modules with learning-based alternatives. A further challenge arises from variability in

the linear adaptive echo cancellation front-end in hybrid approaches. Different studies employ LMS-based, RLS-based, or frequency domain-based adaptive filters approaches such as Multidelay Block Frequency Domain (MDF) [5, 15], each exhibiting different convergence behaviour, latency characteristics, and computational requirements. As a result, improvements attributed to neural enhancement modules may be partially influenced by differences in the underlying echo cancellation stage, complicating fair comparison across systems. Furthermore, many research-oriented systems rely on simplified processing chains or idealized AEC front-ends that differ substantially from those used in commercial products. In practice, production-grade AEC systems are highly optimized for low-latency operation, numerical stability, and computational efficiency. This means that performance gains reported under laboratory conditions may not directly translate to embedded or real-time deployment scenarios.

Among neural enhancement approaches, considerable variation also exists in the choice of learning target. While some methods directly estimate enhanced speech or residual echo signals, others employ mask-based formulations. The Complex Ratio Mask (CRM) preserves both magnitude and phase information while operating on the output of the linear AEC stage, allowing the network to refine an already partially enhanced signal rather than reconstructing the desired signal directly. This structure can improve robustness to acoustic mismatch and changing environmental conditions, making CRM particularly attractive for deployment-oriented hybrid AEC systems where generalization beyond the training data is critical [7]. In contrast, direct signal estimation approaches often place a greater burden on the network to model the complete enhancement process and may be more sensitive to variations between training and deployment conditions. Collectively, these factors make it difficult to isolate the true contribution of neural enhancement components within hybrid AEC systems. Reported performance improvements may be influenced not only by the enhancement model itself, but also by differences in front-end adaptive filtering, deployment assumptions, and experimental design choices.

Therefore, there is a need for a controlled and deployment-oriented evaluation framework in which traditional and hybrid enhancement systems operate on an identical production-grade linear AEC front-end. Such a framework enables the contribution of the neural enhancement stage to be isolated while simultaneously evaluating whether the performance gains reported in the literature translate to realistic real-time processing environments.

Motivated by these considerations, this thesis investigates a CRM-based hybrid AEC architecture designed for low-latency operation and directly compares it against a proprietary traditional speech enhancement pipeline incorporating RES and NS, provided by an industry partner. By retaining an identical linear AEC front-end across both systems, the proposed framework enables a controlled and fair assessment of the practical benefits and trade-offs associated with integrating deep learning-based enhancement into real-world AEC pipelines.

1.6 Contributions of This Thesis

This thesis addresses challenges associated with the fair evaluation and practical deployment of hybrid AEC systems through the development of a deployment-oriented deep learning enhancement framework and a controlled experimental methodology. The main contributions are summarized as follows:

- The development of a hybrid AEC architecture derived from the convolutional recurrent network proposed in [15], redesigned for real-time and deployment-oriented operation. Unlike the original system, which directly estimates the near-end speech signal, the proposed model uses a CRM-based enhancement strategy following linear echo cancellation. The CRM provides a more generalizable and practical learning target by enabling the network to learn a time–frequency transformation of the residual signal rather than directly reconstructing the near-end speech. This allows the learned component to focus on residual echo and noise while the computationally efficient linear AEC handles the primary echo cancellation task.

- The adaptation of the original network architecture to support CRM-based enhancement. This includes replacing the dual real-imaginary decoder structure with a single shared decoder that jointly estimates the complex mask components, removing the auxiliary Voice Activity Detection (VAD) branch, and reducing decoder complexity through the elimination of the final skip connection and corresponding decoder stage. These modifications were introduced to better align the network with the adopted masking objective and to improve suitability for real-time deployment.

- The integration of the proposed neural enhancement model with a production-grade Partitioned Block Frequency-Domain Adaptive Filter (PBFDAF) linear AEC front-end, replacing the idealized adaptive filtering approaches commonly used in prior work. This enables evaluation under conditions that more closely reflect practical embedded and real-time deployment environments.

- The development of a controlled experimental framework in which both the traditional signal processing and deep learning-based systems use the same linear echo cancellation front-end. By keeping the AEC stage identical across all systems, the framework isolates the effect of the subsequent enhancement stage, allowing the conventional AES and NS pipeline to be directly compared with the proposed neural post-filter under otherwise equivalent conditions. This controlled comparison enables the performance differences to be attributed primarily to the choice of enhancement approach rather than differences in the underlying echo cancellation front-end.

- A comprehensive evaluation using standardized ICASSP challenges AEC scenarios, including near-end single talk, far-end single talk, and double-talk conditions. Performance is assessed using both echo suppression and perceptual quality metrics, including ERLE, DNSMOS, and AECMOS, providing insight into the practical benefits and trade-offs of incorporating deep learning-based enhancement into real-world AEC pipelines [1, 16, 17].

1.7 Thesis Organization

The remainder of this thesis is structured as follows:

- **Chapter 2: Background** introduces the theoretical foundations of AEC, classical adaptive filtering techniques, spectral representations, and deep learning-based speech enhancement methods.
- **Chapter 3: Related Works** reviews existing literature in classical AEC, AES, and NS systems, as well as recent deep learning and hybrid approaches, and identifies key limitations motivating this work.
- **Chapter 4: Proposed Hybrid AEC System** presents the architecture of the proposed system, including the PBFDAF front-end, deep learning post-filter, and CRM-based enhancement formulation.
- **Chapter 5: Experimental Setup** describes the datasets, training procedure, evaluation scenarios, and objective metrics used to assess system performance.
- **Chapter 6: Results and Analysis** presents quantitative and qualitative results, including comparisons with a real-world cascaded AEC system, temporal analysis using ERLE, and computational performance evaluation.
- **Chapter 7: Conclusion and Future Work** summarizes the key findings of the thesis and outlines potential directions for future research.

Chapter 2

Background

This chapter provides the theoretical and technical background necessary for the development of the proposed hybrid AEC system. The chapter begins by introducing the acoustic echo path and its effect on hands-free communication systems. After this, since many modern AEC and speech enhancement systems operate in the time-frequency domain with block processing, the chapter also reviews STFT-based processing and block processing. Classical linear acoustic AEC techniques are then reviewed, with emphasis on adaptive filtering methods commonly used for echo estimation and removal.

Following the discussion of linear AEC and STFT-based processing, RES and NS techniques are presented to address the remaining echo components and background noise that persist after linear cancellation.

Finally, the last chapter introduces deep learning approaches for speech enhancement and echo suppression, highlighting the transition from traditional signal processing methods toward data-driven architectures.

Together, these topics establish the foundation required for understanding the proposed hybrid deep learning and adaptive signal processing framework presented in later chapters.

2.1 Acoustic Echo Path

2.1.1 Signal Model

In hands-free communication systems, two users communicate from different physical locations through loudspeakers and microphones. The speech signal arriving from the remote side of the communication link is referred to as the *far-end signal*, while the speech produced locally near the microphone is referred to as the *near-end signal* [1].

At the near-end side, the far-end signal is reproduced through a loudspeaker and propagates through the acoustic environment before being captured by the microphone. This undesired coupling between the loudspeaker and microphone produces an acoustic echo component that is transmitted back to the far-end user, degrading communication quality and speech intelligibility [18].

In AEC systems, the observed microphone signal consists of multiple components, including the desired near-end speech, the far-end echo, and background noise. The acoustic echo signal can be modeled as the convolution between the far-end reference signal and the acoustic impulse response of the environment often referred to as the Room Impulse Response (RIR):

$$y[n] = x[n] * h[n]. \tag{2.1}$$

where $x[n]$ is the far-end signal, $h[n]$ represents the acoustic echo path, and $y[n]$ is the resulting echo signal captured at the microphone. The symbol $*$ denotes the linear convolution operation [19, 1].

Assuming a Finite Impulse Response (FIR) representation of the acoustic path with filter length L, the convolution can be expanded as

$$y[n] = \sum_{k=0}^{L-1} h[k]x[n-k]. \tag{2.2}$$

where $h[k]$ denotes the impulse response coefficients of the acoustic environment [19].

The microphone signal can therefore be expressed as

$$d[n] = y[n] + s[n] + v[n]. \tag{2.3}$$

The signal model described by (2.1)–(2.3) is illustrated in Figure 2.1. The far-end signal propagates through the acoustic echo path to produce the echo component captured at the microphone, where it combines with the desired near-end speech and background noise.

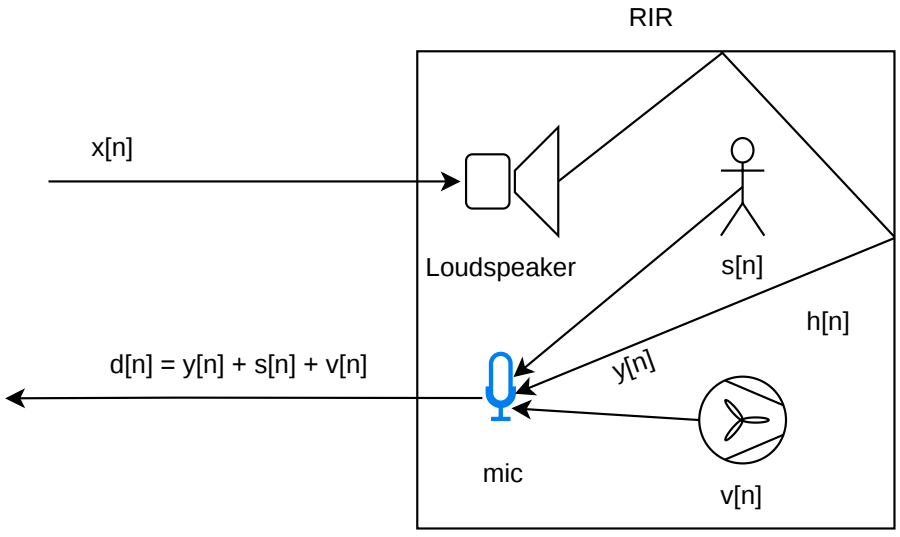


Figure 2.1: Acoustic echo signal model illustrating the formation of the microphone signal from the far-end signal, acoustic echo path, near-end speech, and background noise.

where:

- $d[n]$ is the microphone signal,
- $y[n]$ is the acoustic echo component,
- $s[n]$ is the near-end speech signal,
- and $v[n]$ is additive background noise.

The objective of an AEC system is to estimate the echo component $y[n]$ and remove it from the microphone signal while preserving the desired near-end speech signal [18, 1].

2.1.2 RIR

The acoustic echo path $h[n]$ represents the RIR between the loudspeaker and microphone within the acoustic environment. It models the propagation characteristics of sound waves and captures the effects of:

- Direct-path propagation between the loudspeaker and microphone,
- Reflections from walls, furniture, and surrounding objects,
- Reverberation caused by multiple acoustic reflections.

When the far-end signal is reproduced through the loudspeaker, sound propagates not only along the direct path but also through numerous reflected paths. As a result, the microphone captures multiple delayed and attenuated copies of the original signal [18, 19]. The combination of these propagation effects forms the overall RIR.

In practical environments such as conference rooms, vehicles, and smart speaker systems, the RIR may contain a strong direct-path component followed by a long reverberant tail. The length of the impulse response depends on factors such as room size, surface materials, reverberation time, and sampling frequency [18]. Highly reverberant environments can produce long impulse responses requiring a large number of coefficients for accurate modeling.

Since the acoustic echo path includes both the physical acoustic environment and the electroacoustic characteristics of the loudspeaker and microphone, accurate estimation of $h[n]$ is needed for effective echo cancellation performance.

2.1.3 Time-Varying Nature of the Echo Path

In practical applications, the acoustic echo path is not static but varies over time. Changes in the environment can alter the impulse response and consequently affect the characteristics

of the echo signal.

Common causes of echo path variations include:

- Movement of speakers, microphones, or users,
- Changes in the surrounding acoustic environment,
- Variations in loudspeaker or microphone characteristics,
- Changes in room configuration or object placement.

These variations cause the RIR to become time-varying, meaning that the echo path coefficients change continuously during operation [18, 1]. As a result, AEC systems must continuously adapt their filter coefficients in order to track the changing environment.

The time-varying nature of the echo path represents a major challenge for echo cancellation systems because inaccurate tracking can lead to increased residual echo and degraded speech quality.

2.1.4 Challenges in Echo Modeling

Accurate modeling of the acoustic echo path is challenging due to several practical factors encountered in real-world communication systems. One major difficulty is the potentially long impulse response length in reverberant environments, which increases the number of coefficients that must be estimated by the adaptive filter [18].

The presence of background noise and near-end speech further complicates echo estimation because these signals interfere with the adaptive filtering process. In particular, simultaneous far-end and near-end speech, commonly referred to as double-talk, can significantly degrade adaptation performance if not properly handled [20].

In addition, practical loudspeakers and audio hardware may introduce nonlinear distortions such as clipping, harmonic distortion, and amplifier nonlinearities. Since conventional adaptive filters are primarily designed for linear system modeling, these nonlinear effects are

difficult to estimate accurately and often result in residual echo and noise remaining after cancellation [18].

These challenges limit the performance of classical linear echo cancellation systems and motivate the development of more advanced adaptive filtering, RES, and deep learning-based approaches.

The acoustic echo path therefore forms the foundation of AEC systems, since the objective of the canceller is to estimate and remove the echo component from the microphone signal. However, before discussing AEC in detail, it is important to first introduce the STFT and block processing, as these techniques underpin many modern AEC systems.

2.2 Time-Frequency Processing and Block Processing

2.2.1 STFT

In practical acoustic signal processing systems, signals are commonly analyzed in the time-frequency domain using the STFT [19]. Unlike the conventional Fourier transform, which represents the frequency content of an entire signal globally, the STFT provides a localized spectral representation that captures how the frequency content of a signal evolves over time. This is particularly important for speech signals, which are highly non-stationary and exhibit continuously changing spectral characteristics.

The core idea of the STFT is to divide the input signal into short overlapping segments and compute the Fourier transform independently for each segment. By assuming approximate stationarity within each short analysis window, the signal can be represented as a sequence of localized frequency-domain frames suitable for adaptive filtering and spectral enhancement tasks.

Given a discrete-time signal $x[n]$, the STFT is defined as

$$X[m,k] = \sum_{n=-\infty}^{\infty} x[n]w[n-mH]e^{-j\frac{2\pi}{N}kn}. \tag{2.4}$$

where:

- $X[m,k]$ denotes the complex spectral coefficient at frame index m and discrete frequency bin k,
- $x[n]$ is the input time-domain signal,
- $w[\cdot]$ represents the analysis window function,
- H is the hop size (shift, stride) between consecutive frames,
- and N is the Fast Fourier Transform (FFT) length.

The analysis window isolates a short temporal segment of the signal prior to transformation into the frequency domain. Common window functions include the Hann, Hamming, and rectangular windows, each providing different trade-offs between spectral leakage and frequency resolution [19]. Consecutive frames are typically overlapped in time to reduce discontinuities between adjacent segments and improve reconstruction quality.

The STFT representation is widely used in AEC, AES, and NS systems because it enables localized analysis of speech, echo, and noise across both time and frequency dimensions [3]. These frameworks are further discussed in subsequent sections.

2.2.2 Magnitude and Phase Representation

The resulting STFT transform produces a complex-valued time-frequency representation. Each spectral coefficient can be expressed in polar form as

$$X[m,k] = |X[m,k]|e^{j\angle X[m,k]}. \tag{2.5}$$

where $|X[m,k]|$ denotes the magnitude spectrum and $\angle X[m,k]$ represents the phase spectrum corresponding to frame index m and frequency bin k.

The magnitude component describes the distribution of spectral energy across frequency bins and is strongly related to the perceptual characteristics of speech, such as formants and harmonic structure. Due to this fact, many classical speech enhancement and adaptive filtering algorithms primarily focus on modifying the magnitude spectrum [3]. In these approaches, time-frequency dependent suppression gains or masks are applied directly to the spectral magnitude, and the enhanced time-domain signal is subsequently reconstructed using the unmodified noisy phase information.

Historically, phase information was often considered perceptually less important than magnitude, particularly in early speech enhancement systems operating at moderate Signal-to-Noise Ratio (SNR)s [3].

Although this approximation simplifies processing considerably, it introduces limitations because phase contains important temporal and structural information about the underlying speech signal. In particular, accurate phase representation becomes increasingly important for high-quality speech reconstruction in highly degraded acoustic conditions [21]. Errors in phase can lead to temporal smearing, reduced speech naturalness, and degraded perceptual quality even when the magnitude spectrum is accurately estimated.

Modern speech enhancement systems therefore increasingly incorporate explicit phase-aware processing techniques. Recent approaches may estimate complex-valued spectral masks, directly predict real and imaginary spectral components, or jointly enhance both magnitude and phase information [21, 15]. These methods aim to improve reconstruction quality beyond what is achievable through magnitude-only enhancement.

2.2.3 Block Processing for Real-Time Systems

In real-world AEC systems, processing is performed in a block-wise manner to satisfy latency constraints. Instead of processing the entire signal at once, the input is divided into

overlapping frames processed sequentially.

This block processing framework introduces a trade-off between:

- Latency (shorter blocks reduce delay)
- Frequency resolution (longer blocks improve spectral accuracy)
- Computational cost

Overlap-add or overlap-save techniques are typically used to reconstruct the time-domain signal from processed frames.

For real-time applications such as hands-free communication or hearing aids, block size selection is critical to balancing performance and latency requirements.

The time-frequency and block-based processing techniques discussed in this section form the computational foundation of many classical speech enhancement systems, including AEC, RES, and NS algorithms. With this foundation in place, the next section introduces linear AEC algorithms as a first step toward more advanced adaptive and data-driven approaches.

2.3 Linear AEC Algorithms

2.3.1 Overview of Linear AEC

Linear AEC systems aim to estimate and remove the acoustic echo component from the microphone signal using an adaptive filter that models the acoustic echo path [18, 1]. Since the acoustic echo path can be represented as a linear FIR system, the echo signal may be estimated by filtering the far-end reference signal through an adaptive filter whose coefficients approximate the unknown acoustic RIR [19, 1].

Conceptually, the adaptive filter attempts to recreate the RIR between the loudspeaker and microphone. By continuously adapting its coefficients, the filter learns an approximation of these propagation characteristics such that its output closely matches the actual echo observed at the microphone [1].

If the adaptive filter accurately converges to the true acoustic impulse response, the generated echo estimate becomes nearly identical to the real echo component present in the microphone signal. Subtracting this estimate from the microphone signal therefore suppresses the far-end echo while preserving the desired near-end speech [18]. In this sense, the adaptive filter acts as an online estimator of the acoustic environment, continuously updating its coefficients to track changes in the room acoustics and loudspeaker-to-microphone propagation path [1].

Let $\hat{h}[n]$ denote the adaptive estimate of the acoustic echo path. The estimated echo signal $\hat{y}[n]$ can then be expressed as

$$\hat{y}[n] = x[n] * \hat{h}[n]. \tag{2.6}$$

or equivalently,

$$\hat{y}[n] = \sum_{k=0}^{L-1} \hat{h}[k]x[n-k]. \tag{2.7}$$

where $\hat{h}[k]$ represents the estimated adaptive filter coefficients and L is the filter length [1, 19].

The estimated echo signal is subtracted from the microphone signal to generate the residual error signal

$$e[n] = d[n] - \hat{y}[n]. \tag{2.8}$$

This residual error signal serves as the objective function used to adapt the filter coefficients. Most adaptive algorithms seek to minimize the Mean Squared Error (MSE) between the microphone signal and the estimated echo by iteratively adjusting the filter coefficients in the direction that reduces the error [1]. This optimization process, which can be viewed as a form of gradient-based adaptation, enables the filter to progressively improve the echo estimate while tracking changes in the acoustic environment. As a result, the adaptive filter

continuously refines its estimate of the acoustic path and reduces residual echo over time.

The operation of a linear AEC is illustrated in Figure 2.2. The far-end reference signal is processed by an adaptive filter to generate an estimate of the acoustic echo signal. This estimate is subsequently subtracted from the microphone signal to produce the residual error signal. The residual error is then used by the adaptation algorithm to update the filter coefficients and continuously track changes in the acoustic environment.

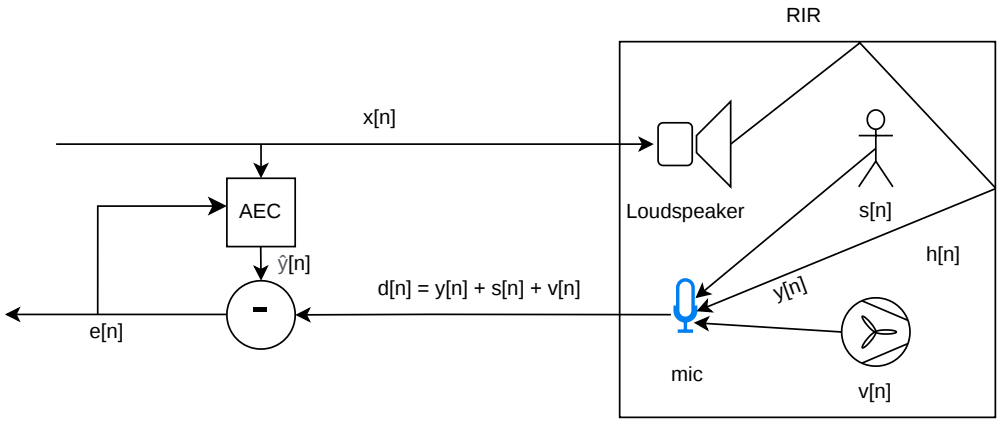


Figure 2.2: Block diagram of a linear AEC system. The adaptive filter estimates the echo signal using the far-end reference signal, and the resulting residual error signal is used to adapt the filter coefficients.

Ideally, the residual signal contains only the near-end speech and background noise components. In practice, however, some residual echo often remains due to imperfect echo path estimation, model mismatch, environmental changes, and nonlinear distortions [18].

Linear AEC systems form the foundation of many practical echo cancellation architectures due to their relatively low computational complexity and effectiveness in modeling linear acoustic propagation [18].

Frequency-Domain Adaptive Filtering

As acoustic impulse responses become longer in reverberant environments, time-domain adaptive filters become increasingly computationally expensive. In a time-domain implementation, the echo signal is generated through a convolution operation between the far-end

signal and a long adaptive filter. For an impulse response of length L, this requires L multiplications and additions per output sample, making real-time processing computationally demanding for long echo paths [1, 2].

To address this limitation, frequency-domain adaptive filtering approaches were developed for AEC [2]. These methods exploit the convolution theorem, whereby convolution in the time domain is equivalent to multiplication in the frequency domain [19]. By transforming signals into the frequency domain, the convolution operation can be implemented using efficient spectral-domain multiplication, substantially reducing computational complexity for long adaptive filters [2, 19].

More specifically, instead of directly computing the estimated echo signal as

$$\hat{y}[n] = x[n] * h[n], \tag{2.9}$$

the filtering operation can be performed in the frequency domain as

$$\hat{Y}[k, m] = X[k, m] \cdot H[k, m]. \tag{2.10}$$

where $X[k, m]$, $H[k, m]$, and $\hat{Y}[k, m]$ denote the frequency-domain representations of the far-end signal, adaptive filter, and estimated echo, respectively.

For the long impulse responses commonly encountered in AEC, further computational savings can be achieved by partitioning the adaptive filter into multiple shorter segments. This forms the basis of partitioned block frequency-domain adaptive filtering (PBFDAF), where individual filter partitions are processed independently in the frequency domain and combined to produce the overall echo estimate [22].

A widely used realization of this concept is the multidelay filter (MDF) algorithm, which enables efficient adaptive filtering of long acoustic impulse responses [23].

Although frequency-domain adaptive filtering significantly improves computational efficiency compared to time-domain approaches, linear AEC systems are generally unable to

completely eliminate acoustic echo in practical environments. Residual echo often remains due to model mismatch, finite filter length, time-varying acoustic paths, double-talk conditions, and nonlinear distortions that cannot be fully captured by linear adaptive filters. Therefore, additional processing stages are typically employed to further suppress the remaining echo components. This motivates the use of RES, which is introduced in the following section.

2.4 RES

2.4.1 Overview

Linear AEC systems are inherently limited in their ability to fully suppress echo components in practical acoustic environments. As a result, a non-negligible residual error remains after adaptive filtering. This residual component, referred to as the *residual echo*, arises from model mismatch, finite filter length, time-varying acoustic paths, and nonlinear distortions that cannot be captured by linear filtering approaches [24, 1]. This residual signal often overlaps spectrally with near-end speech and background noise, making it difficult to suppress using linear filtering alone without introducing speech distortion [24].

RES, also referred to as AES, is therefore introduced as a post-processing stage that further attenuates the remaining echo components after linear cancellation. RES methods typically operate as signal enhancement techniques in the time-frequency domain. They aim to estimate time-frequency masks or gains that selectively reduce residual echo energy while preserving near-end speech intelligibility [25].

In modern speech communication systems, RES is commonly integrated as part of a multi-stage pipeline following linear AEC. The linear stage provides a physically motivated estimate of the echo component, while RES refines the output by addressing the limitations of linear modeling in real-world acoustic conditions. This separation allows the overall system to achieve improved echo attenuation without excessively increasing the complexity of the

adaptive filtering stage [24, 25].

2.4.2 Spectral Domain Formulation

In most practical RES systems, processing is performed in the STFT domain. This representation is particularly well suited to speech enhancement tasks because it provides a joint time-frequency decomposition of the signal, allowing different spectral components to be processed independently.

After linear AEC, the microphone signal can be expressed in the spectral domain as a sum of three main components:

$$D[k, f] = S[k, f] + Y_r[k, f] + V[k, f]. \tag{2.11}$$

where $D[k, f]$ denotes the microphone signal in the time-frequency domain, $S[k, f]$ is the desired near-end speech component, $Y_r[k, f]$ represents the residual echo remaining after linear cancellation, and $V[k, f]$ corresponds to additive background noise.

Unlike the linear AEC stage, which attempts to explicitly model the acoustic echo path, residual echo suppression operates directly on this mixed observation. The objective is to estimate a time-frequency dependent gain function $G[k, f]$ that selectively attenuates undesired components while preserving near-end speech content. The enhanced speech estimate is then obtained by applying this gain on the microphone signal:

$$\hat{S}[k, f] = G[k, f] \cdot D[k, f]. \tag{2.12}$$

This formulation provides a flexible framework in which suppression can be adapted across both time and frequency dimensions. Regions dominated by residual echo or noise can be attenuated more strongly, while speech-dominant regions are preserved through higher gain values. As a result, RES methods naturally extend classical filtering approaches into adaptive spectral weighting techniques, forming the basis for both traditional statistical estimators

and modern learning-based enhancement models [24, 25].

2.4.3 Wiener Filtering Approach

A classical solution to RES is Wiener filtering, which provides an optimal linear estimator in the MSE sense under the assumption that the underlying signal components are stationary and statistically independent [1]. In the context of RES, the goal is to estimate the clean near-end speech component from a corrupted observation consisting of speech, residual echo, and background noise.

Under these assumptions, the optimal frequency-dependent gain function can be expressed in terms of the power spectral densities of the constituent signals. The Wiener gain is defined as:

$$G[k,f] = \frac{\Phi_{SS}[k,f]}{\Phi_{SS}[k,f] + \Phi_{YY}[k,f] + \Phi_{VV}[k,f]}. \tag{2.13}$$

where $\Phi_{SS}[k,f]$ denotes the Power Spectral Density (PSD) of the desired near-end speech, $\Phi_{YY}[k,f]$ represents the PSD of the residual echo component, and $\Phi_{VV}[k,f]$ corresponds to the PSD of the background noise [25].

This formulation highlights that the suppression behavior of the Wiener filter is governed entirely by the relative energy distribution between the desired signal and the interfering components. When the speech power dominates, the gain approaches one, preserving the signal. Conversely, when residual echo or noise dominates, the gain is reduced, resulting in stronger attenuation.

In practical systems, the individual PSDs are not directly observable and must instead be estimated from the available microphone signal and prior echo cancellation stages. This requirement leads to a range of practical approximations of the Wiener solution, which form the basis for many classical and modern speech enhancement algorithms.

2.4.4 Spectral Gain-Based Methods

The impracticality of directly computing the Wiener solution in real-world RES systems has led to a broad class of spectral gain-based methods. These approaches operate by estimating time-varying gains in the spectral domain using observable quantities derived from the microphone signal and intermediate echo cancellation outputs. Rather than relying on explicit knowledge of the underlying PSDs, these methods construct approximations that enable adaptive suppression of residual echo under non-stationary conditions [25, 26].

This leads to a broad class of spectral gain-based methods, which implement RES by computing a frequency-dependent attenuation function based on estimated signal-to-interference measures. These methods are widely used in practical speech enhancement systems due to their balance between computational efficiency and perceptual performance [25].

A classical approach within this framework is spectral subtraction, where an estimate of the interference spectrum (residual echo and noise) is removed from the observed microphone spectrum in order to recover an approximation of the clean speech signal [27]. The method operates under the assumption that the noisy observation can be decomposed as the sum of a desired speech component and an additive interference component in the time-frequency domain.

Let the observed magnitude spectrum be expressed as

$$|D[k,f]| = |S[k,f]| + |I[k,f]|. \tag{2.14}$$

where $|I[k,f]|$ represents the combined contribution of residual echo and background noise. In practice, $|I[k,f]|$ is not directly known and must be estimated, typically using noise tracking or minimum statistics approaches [26].

Spectral subtraction then estimates the clean speech magnitude spectrum as

$$|\hat{S}[k,f]| = \max\left(|D[k,f]| - |\hat{I}[k,f]|,\, 0\right). \tag{2.15}$$

where $|\hat{I}[k, f]|$ is the estimated interference magnitude. The $\max(\cdot, 0)$ operation is used to prevent negative magnitude estimates, which would otherwise be physically meaningless.

The corresponding phase is typically assumed to remain unchanged from the noisy observation, leading to the reconstructed complex spectrum

$$\hat{S}[k, f] = |\hat{S}[k, f]|e^{j\angle D[k,f]}. \tag{2.16}$$

While spectral subtraction is computationally efficient and historically significant in speech enhancement, it is highly sensitive to errors in the interference estimate. Inaccurate tracking of residual echo or noise leads to over-subtraction or under-subtraction, which can introduce perceptually disturbing artifacts commonly referred to as *musical noise* [27, 28]. These artifacts arise from rapid, random fluctuations across individual frequency bins caused by independent bin-wise processing.

More advanced approaches rely on explicit estimation of the a priori and a posteriori SNRs, which provide more stable statistical measures for gain computation compared to direct spectral subtraction. Rather than attempting to subtract an estimated interference spectrum, these methods formulate suppression as an estimation problem based on the relative power of the desired signal and the observed noisy signal in each time-frequency bin [25].

The a posteriori SNR is defined as the instantaneous ratio between the observed signal power and the estimated interference power:

$$\gamma[k, f] = \frac{|D[k, f]|^2}{\lambda_I[k, f]}. \tag{2.17}$$

where $\lambda_I[k, f]$ denotes the PSD of the interference, which includes both residual echo and background noise.

In contrast, the a priori SNR represents an estimate of the clean speech power relative

to the interference and is typically computed recursively to provide temporal smoothing:

$$\xi[k, f] = \frac{\lambda_S[k, f]}{\lambda_I[k, f]}. \tag{2.18}$$

where $\lambda_S[k, f]$ is the estimated clean speech power. Since $\lambda_S[k, f]$ is not directly observable, it is commonly estimated using decision-directed approaches that combine the previous enhanced speech estimate with the current observation [25].

Using these SNR estimates, the Wiener-style suppression gain can be expressed as a function of the a priori SNR:

$$G[k, f] = \frac{\xi[k, f]}{1 + \xi[k, f]}. \tag{2.19}$$

This formulation provides a smoother and more stable gain compared to spectral subtraction because it avoids direct bin-wise subtraction and instead relies on statistically averaged power estimates. As a result, abrupt fluctuations in the gain function are reduced, leading to improved perceptual quality and reduced musical noise artifacts.

In practice, the combination of a posteriori and a priori SNR estimates allows the system to balance responsiveness and stability. The a posteriori SNR captures instantaneous signal conditions, while the a priori estimate provides temporal smoothing across frames. This balance is important in RES, where interference statistics can vary rapidly due to double-talk and time-varying acoustic conditions.

A further extension of these techniques is the use of mask-based suppression functions, where the gain is interpreted as a soft time-frequency mask applied to the noisy spectrum. These masks can be derived from statistical models or learned directly from data, and they allow more flexible control over the trade-off between interference suppression and speech preservation [29]. In this formulation, frequency bins dominated by residual echo or noise are strongly attenuated, while speech-dominant regions are preserved through higher gain values.

Across all these approaches, the central idea remains consistent: suppression is achieved

by estimating a time-frequency dependent gain that reflects the local signal-to-interference ratio.

However, regardless of the specific residual suppression approach employed, the output signal often still contains a non-negligible amount of residual noise and distortion. This motivates the use of dedicated NS techniques within the broader speech enhancement framework. These approaches operate primarily in the spectral domain and focus on estimating and attenuating noise components using statistical models of speech and noise. Compared to basic residual suppression approaches, NS techniques typically provide more structured estimation strategies and improved robustness through explicit modelling of noise statistics, contributing to stable and perceptually meaningful enhancement. The following section introduces classical NS methods and their role in speech enhancement pipelines.

2.5 Noise Suppression (NS)

2.5.1 Overview

In practical acoustic communication systems, the microphone signal is affected not only by acoustic echo components but also by various forms of additive background noise. Common noise sources include ambient environmental sounds, road noise in automotive environments, fan noise, air conditioning systems, keyboard noise, and microphone self-noise [28]. These unwanted components can significantly degrade speech intelligibility, listening comfort, and the overall quality of hands-free communication systems.

NS aims to attenuate these unwanted noise components while preserving the desired speech signal as much as possible. Similar to RES, most classical NS systems operate in the time-frequency domain and apply frequency-dependent gain functions to selectively reduce interference energy [28, 27].

Unlike residual echo suppression, which targets structured interference correlated with the far-end reference signal, noise suppression typically deals with stochastic and largely

uncorrelated interference signals. As a result, NS algorithms generally rely on statistical estimation of noise characteristics rather than explicit acoustic path modeling.

A major challenge in noise suppression is achieving an appropriate balance between interference attenuation and speech preservation. Excessive suppression may introduce speech distortion and perceptual artifacts, while insufficient suppression leaves residual noise that degrades speech quality [28]. This is why practical NS systems are designed not only to maximize noise reduction, but also to maintain natural speech characteristics and minimize perceptual degradation.

2.5.2 Classical Noise Suppression Methods

Classical NS methods form a foundational class of speech enhancement techniques that operate primarily in the time-frequency domain. While conceptually similar to residual suppression approaches in that they rely on frequency-dependent attenuation, NS methods are typically formulated more generally for the removal of background noise and other additive interferences, rather than being tied to a specific upstream processing stage [28].

Many classical NS algorithms employ spectral gain estimation techniques similar to those introduced in the previous section on RES. Common approaches include spectral subtraction, Wiener filtering, Minimum Mean Squared Error (MMSE) estimators, and minimum-statistics-based noise estimation methods [27, 25, 28, 26, 30]

Despite their effectiveness, classical NS methods generally rely on assumptions of local stationarity and accurate estimation of noise PSDs. In practical scenarios involving rapidly changing noise conditions, overlapping speakers, or highly non-stationary interference, these assumptions are often violated, leading to reduced suppression performance and perceptual degradation.

Modern practical systems therefore often employ more advanced noise estimation strategies that combine multiple statistical cues and adaptive control mechanisms to improve robustness under dynamic acoustic conditions. The baseline system used in this work em-

ploys such a multi-factor noise estimation framework, which is discussed in a section 4.5 [30, 31].

2.5.3 Interaction with Residual Echo Suppression

In practical AEC systems, NS operates alongside RES as part of a multi-stage speech enhancement pipeline. After the linear AEC stage removes the dominant echo component, the remaining signal still contains residual echo, background noise, and desired near-end speech. Consequently, both RES and NS must process a shared residual signal that contains multiple overlapping interference sources [4].

A major difficulty arises because residual echo and background noise frequently overlap within the same time-frequency regions and can exhibit similar spectral characteristics. In highly reverberant environments, residual echo may appear noise-like due to diffuse late reflections and nonlinear distortion components, making it difficult to distinguish from true background noise using purely statistical methods [4]. As a result, the operation of NS and RES becomes strongly coupled in practical enhancement systems.

This interaction introduces several important challenges:

- Over-suppression of desired speech components when residual echo or noise is inaccurately estimated,
- Reduced robustness of noise estimation algorithms in the presence of strong residual echo leakage,
- Difficulty balancing interference attenuation against preservation of speech quality and intelligibility,
- Increased susceptibility to speech distortion and perceptual artifacts during double-talk conditions.

Because both suppression stages operate on partially overlapping interference components, independent optimization of RES and NS can lead to conflicting gain decisions across time-frequency bins. Excessive attenuation may improve interference reduction while simultaneously degrading speech naturalness, whereas conservative suppression preserves speech quality at the expense of increased residual interference.

For these reasons, modern speech enhancement systems increasingly treat echo suppression and noise suppression as jointly related problems rather than completely independent processing stages [4]. Practical systems often share statistical estimators, jointly estimate suppression masks, or apply coordinated adaptive control mechanisms across both stages in order to improve overall enhancement performance.

This close relationship between echo suppression and noise suppression has motivated the development of integrated deep learning-based speech enhancement systems, where a single model can jointly learn the complex interactions between speech, residual echo, and background noise directly from data. The following section therefore introduces data-driven deep learning approaches that aim to address the limitations of traditional processing pipelines.

2.6 Deep Learning for Speech Enhancement

2.6.1 Overview

Deep learning has become a dominant approach in modern speech enhancement systems, significantly improving performance over many traditional signal processing methods in challenging acoustic environments [29]. Rather than relying solely on hand-crafted statistical models or analytically derived suppression rules, deep neural networks learn nonlinear mappings directly from corrupted speech signals to enhanced speech representations using large-scale training data.

Figure 2.3 illustrates a generic deep learning-based speech enhancement framework operating on time-frequency representations, which serves as a common abstraction for a wide

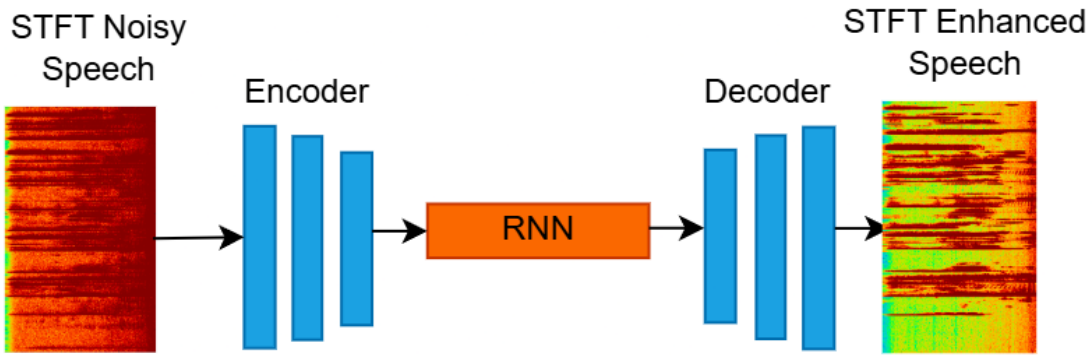


Figure 2.3: Generic deep learning-based speech enhancement pipeline operating in the time-frequency domain. A corrupted speech signal is transformed into a spectral representation using the STFT, processed by an encoder–bottleneck–decoder architecture incorporating temporal modeling RNN, and mapped to an enhanced speech spectrogram.

range of neural architectures discussed in this section.

In AEC systems, deep learning is typically incorporated in one of two ways: either as an end-to-end replacement of the full pipeline or as part of a hybrid architecture built on top of a conventional linear AEC stage [32, 15]. In the hybrid case, a linear AEC module is first used to perform coarse echo cancellation, after which a deep learning model is applied to further suppress residual echo, background noise, reverberation, and other unmodelled distortions in the time-frequency domain. In fully deep learning-based approaches, both the linear cancellation and post-processing stages are replaced by a single model that directly estimates the desired clean speech signal. Both of these approaches are attractive because they can learn complex relationships between speech and interference signals that are difficult to represent using classical statistical assumptions alone.

Most deep learning-based enhancement systems operate on spectral representations derived from the STFT, where neural networks estimate suppression masks, spectral gains, or enhanced speech spectra from corrupted microphone observations. Depending on the formulation, the network may operate on magnitude spectra or complex-valued spectral representations [7].

A major advantage of deep learning approaches is their ability to model highly non-stationary acoustic conditions. Unlike traditional suppression methods that depend on ex-

plicit estimation of signal statistics such as noise PSDs or SNRs, neural networks can implicitly learn robust enhancement behavior directly from data across a wide range of acoustic scenarios [29].

Additionally, deep neural networks can incorporate both local spectral structure and long-term temporal dependencies through architectures such as convolutional and recurrent neural networks. This enables improved modeling of speech continuity, reverberation tails, residual echo patterns, and dynamically changing noise conditions compared to many classical enhancement approaches.

The following section does not attempt to comprehensively review individual deep learning architectures, masking strategies, or complex-domain enhancement methods, as these topics are examined in detail in Chapter 3. Instead, this section provides the general background and motivation for integrating deep learning within modern speech enhancement and hybrid AEC systems.

2.6.2 CNN Based Processing

CNNs are widely used in speech enhancement and acoustic interference suppression due to their ability to learn localized patterns directly from time-frequency representations [33, 34]. In these systems, convolution operations apply small learnable kernels, which are matrices of trainable weights, across neighboring regions of the input representation. As the kernel slides over the time-frequency representation, element-wise multiplications and summations are performed to extract localized spectral and temporal patterns associated with speech, noise, and residual echo components.

Unlike fully connected architectures, convolutional layers process only small receptive fields of the input at a time while reusing the same kernel weights across the entire representation. This weight sharing mechanism significantly reduces the number of trainable parameters and improves generalization performance [35]. Such processing is particularly well suited for acoustic signals, where strong correlations often exist between adjacent time

frames and neighboring frequency bins.

In speech enhancement applications, convolutional kernels slide across spectral inputs to identify localized acoustic features such as harmonic structures, formant transitions, transient events, and residual echo patterns. Early convolutional layers, which operate directly on the input spectral representation, primarily capture local time-frequency patterns such as harmonic structure, spectral edges, and formant-related features. As additional convolutional layers are stacked deeper within the network, the effective receptive field and representational capacity increase, enabling the network to model progressively more complex acoustic structures and longer-range dependencies associated with speech, noise, and residual echo components.

The localized nature of convolutional processing also provides robustness to small spectral variations and non-stationary distortions, making CNN-based systems effective for suppressing background noise and structured interference such as residual acoustic echo. As a result, convolutional architectures are widely incorporated into modern NS, AES, and hybrid AEC systems [10].

However, standard convolutional processing is inherently limited to the receptive field defined by the kernel size and network depth. Although deeper architectures can expand the effective receptive field, modeling long-range temporal dependencies such as extended reverberation tails, slowly varying echo paths, and long-duration speech context remains challenging using convolution alone. Therefore, many modern speech enhancement systems integrate convolutional processing with encoder-decoder frameworks and/or recurrent temporal modeling architectures to better capture both local spectral structure and long-term temporal dependencies.

2.6.3 Encoder-Decoder Architectures

Encoder-decoder architectures are a popular choice in speech enhancement systems as a structured framework for learning transformations between corrupted input representations

and enhanced speech outputs [10]. These architectures are particularly effective in acoustic signal processing due to their ability to progressively compress and reconstruct hierarchical feature representations.

In a typical encoder-decoder architecture, the encoder progressively compresses the input representation into a smaller feature representation using a sequence of convolutional layers [10, 36]. As the signal propagates through the encoder, the spatial dimensions of the feature maps are gradually reduced using operations such as strided convolutions or pooling, while the number of feature channels is typically increased. This process enlarges the receptive field of the network and allows deeper layers to capture increasingly abstract acoustic representations associated with speech structure, residual echo, and noise characteristics.

The compressed representation produced by the encoder retains the most important information needed for reconstructing the speech signal while reducing the influence of interference components such as noise and residual echo. The decoder then performs the inverse operation by progressively reconstructing the enhanced representation through upsampling and convolutional processing, gradually restoring the original time-frequency resolution.

Compared to purely convolutional feature extractors operating at a fixed resolution, encoder-decoder architectures provide a more effective mechanism for aggregating both local spectral detail and broader contextual information across the input representation [10].

A key design element in many encoder-decoder systems is the use of skip connections between corresponding encoding and decoding layers. These connections allow fine-grained structural details that may be lost during the compression stage to be directly transferred to the reconstruction stage. This improves the preservation of harmonics and transient structure, which are often degraded in noisy or reverberant conditions [37].

In speech enhancement and acoustic interference suppression, encoder-decoder architectures are particularly effective for learning mappings that jointly suppress background noise and residual echo while preserving speech intelligibility. Their structured design also makes them well suited for integration with recurrent layers in more advanced hybrid systems cov-

ered next.

2.6.4 Sequence Modeling with Recurrent Networks

Speech signals exhibit strong temporal structure, where information is distributed across both short-term phonetic events and longer-term contextual dependencies. As a result, effective speech enhancement systems often benefit from mechanisms capable of modeling temporal correlations beyond local spectral context [38, 39].

RNNs provide a natural framework for sequence modeling by processing sequential data one time step at a time while maintaining an internal hidden state that evolves throughout the sequence [38]. At each time step, the network receives both the current input frame and the hidden state from the previous time step. The hidden state acts as a form of memory that carries contextual information from earlier portions of the sequence, allowing the network to model temporal dependencies in signals such as speech.

In speech enhancement applications, this sequential processing enables recurrent networks to capture how spectral characteristics evolve over time rather than treating each frame independently. As speech and interference signals often exhibit strong temporal continuity, RNN-based systems can better model slowly varying noise, reverberation effects, and residual echo components that extend across multiple frames.

Among recurrent architectures, Long Short-Term Memory (LSTM) networks are widely used due to their ability to address the vanishing gradient problem commonly encountered in standard RNNs [40]. In conventional recurrent networks, information from earlier time steps can gradually diminish during training as gradients become increasingly small when propagated across long sequences. This makes it difficult for the network to learn long-range temporal relationships.

LSTM networks mitigate the vanishing gradient problem through an internal memory cell and explicit gating mechanisms that regulate how information is stored, updated, and discarded over time. These gates maintain a stable gradient flow during backpropagation,

allowing the network to retain relevant contextual information across extended time intervals. For these reasons, LSTM-based architectures are better suited for modeling the long-term temporal dependencies associated with speech continuity, late reverberation tails, and slowly varying interference components in complex acoustic environments [41].

An important consideration when training recurrent architectures is how the hidden state is handled between training sequences. In *stateless* training, the hidden state is reset at the beginning of each training sample, causing each sequence segment to be processed independently. This approach simplifies optimization and is commonly used when training on fixed-length speech segments, as it allows samples within a batch to be processed in parallel [38].

In contrast, *stateful* training preserves the hidden state across consecutive segments, enabling the network to maintain contextual information beyond the boundaries of an individual training sequence. This can be advantageous when modeling long-term temporal dependencies, such as long echo tails and time varying noise. However, stateful training introduces additional implementation complexity, requires careful management of sequence ordering, and may increase the risk of error accumulation if the hidden state becomes corrupted over time. Therefore, stateful approaches are often most beneficial in applications where capturing extended temporal context is valuable, particularly when using recurrent architectures such as LSTM-based and frequency-time recurrent networks.

Frequency-time recurrent structures are particularly well suited to leveraging extended temporal context due to their ability to model dependencies across both the temporal and spectral dimensions. This enhanced modeling capability improves the representational capacity of recurrent networks for complex acoustic signals, making them highly effective for speech enhancement and AEC tasks [42].

2.6.5 Mask-Based and Spectral Mapping Approaches

Deep learning-based speech enhancement systems typically adopt one of two main formulation paradigms: mask-based estimation or direct spectral mapping. These two approaches differ in how the enhanced speech signal is represented and reconstructed from the observed microphone input [29].

In mask-based approaches, the neural network estimates a time-frequency gain function that is multiplied to the observed signal representation. The enhancement process can be expressed as:

$$\hat{S}[k, f] = M[k, f] \cdot D[k, f]. \tag{2.20}$$

where $D[k, f]$ is the observed microphone signal in the time-frequency domain, $M[k, f]$ is the estimated mask, and $\hat{S}[k, f]$ is the enhanced speech estimate. In this formulation, the network does not directly generate speech but instead learns a weighting function that selectively attenuates undesired components such as noise and residual echo while preserving speech-dominant regions.

This multiplicative structure constrains the output space of the model, which often improves stability and reduces the likelihood of generating unrealistic spectral artifacts. As a result, mask-based methods are widely used in practical speech enhancement systems due to their robustness and interpretability [21].

In contrast, direct spectral mapping approaches learn a direct transformation from the corrupted input representation to the clean speech spectrum:

$$\hat{S}(k, f) = F_\theta(D(k, f)). \tag{2.21}$$

where $F_\theta(\cdot)$ represents a learned nonlinear mapping parameterized by a neural network. Unlike mask-based methods, this formulation does not explicitly constrain the output to be a filtered version of the input. Instead, the model learns to directly estimate the target

speech representation.

While this increases modeling flexibility and allows the network to potentially correct both magnitude and phase distortions more freely, it also places a greater burden on the learning process, as the model must implicitly learn signal structure, suppression behavior, and reconstruction consistency simultaneously [43, 10].

In practical speech enhancement and acoustic interference suppression systems, mask-based approaches are often preferred in scenarios where stability and interpretability are prioritized, while direct mapping approaches are used when higher modeling flexibility is required. Both formulations are commonly integrated into larger architectures depending on the target application and system constraints.

The methods described in this chapter provide the foundational concepts underlying modern speech enhancement systems, including convolutional feature extraction, encoder-decoder reconstruction frameworks, temporal sequence modeling, and spectral masking formulations. Together, these components form the basis of most contemporary neural architectures used in AEC, AES, and NS systems.

While this section focused on the underlying principles and functional roles of these techniques, the following chapter presents a detailed review of existing research contributions. It examines how these building blocks have been combined in prior work, with particular emphasis on hybrid systems that integrate classical signal processing methods with deep learning-based enhancement models.

Chapter 3

Related Works

This chapter presents a review of existing literature in AEC, AES, and NS systems, as well as more recent deep learning-based and hybrid approaches. The objective of this review is to establish the progression of methodological developments in the field and to identify the limitations of existing approaches that motivate the proposed work.

Rather than providing an exhaustive historical overview, the discussion is organized around three key conceptual areas that reflect the evolution of problem formulations and solution strategies in this domain.

1. **Classical AEC, AES, and NS Methods:** This section reviews traditional signal processing approaches based on adaptive filtering and statistical estimation techniques, including LMS-based echo cancellation and Wiener filtering-based echo and noise suppression.

2. **Deep Learning-Based Speech Enhancement and Echo Suppression:** This section examines approaches that replace or augment classical signal processing models using neural networks operating in the time-frequency or time-domain representations.

3. **Research Gap:** This section discusses the research gap found within the literature review that this paper is targeting.

The organization of this chapter reflects a transition from traditional digital signal processing methods to data-driven and hybrid frameworks, which forms the basis for the system design presented in the subsequent chapters.

3.1 Classical Formulations of AEC, AES, and NS

This section reviews classical approaches to AEC, AES, and NS from a signal processing perspective. The focus is on how these methods formulate the enhancement problem under linear and statistical assumptions, which forms the basis for later comparisons with data-driven and hybrid approaches.

3.1.1 Classical AEC

Classical AEC methods are commonly implemented using adaptive filtering techniques that estimate the acoustic echo path from a far-end reference signal. Early systems relied heavily on LMS and Normalized Least Mean Squares (NLMS) algorithms due to their simplicity and computational efficiency, while RLS-based approaches were investigated to improve convergence and tracking performance in dynamically changing environments [44, 45, 1].

These adaptive filtering methods aim to minimize the error between the microphone signal and the estimated echo component in an iterative manner, allowing the filter coefficients to continuously adapt to variations in the acoustic environment. A comprehensive overview of adaptive filtering techniques for AEC can be found in [1].

As acoustic impulse responses become longer in reverberant environments, conventional time-domain adaptive filtering approaches become increasingly computationally demanding. To address this issue, frequency-domain adaptive filtering techniques were introduced, where convolution operations are efficiently performed using block-based Fourier domain processing [2]. Among these approaches, the PBFDAF algorithm has become widely adopted for practical AEC applications due to its improved computational efficiency and suitability for

long acoustic impulse responses [23, 46].

Despite their widespread use, classical AEC methods rely on the assumption of a predominantly linear acoustic environment and often require careful parameter tuning to maintain stability and convergence in real-world scenarios. Their performance can also degrade significantly under challenging conditions such as double-talk, where near-end and far-end speech are simultaneously active, or in the presence of nonlinear loudspeaker distortions and rapidly changing acoustic paths. Furthermore, while adaptive filters can effectively suppress the dominant linear echo component, residual echo and background noise often remain after cancellation. As a result, complementary approaches for residual echo suppression and noise reduction are commonly employed in the time-frequency domain under statistical estimation frameworks. These methods are generally referred to as AES and NS techniques.

3.1.2 Classical NS and AES

Following linear AEC, RES and NS modules are commonly employed to further enhance speech quality. Early approaches were predominantly based on statistical spectral estimation techniques, including Wiener filtering, MMSE estimators, and spectral subtraction methods [25, 47].

As practical deployment scenarios became increasingly non-stationary, more sophisticated approaches focused on improved estimation of interference PSDs through techniques such as minimum statistics and smoothed minimum statistics noise tracking [47, 26].

These methods became widely adopted in real-time communication systems due to their low computational complexity and straightforward implementation. In conventional processing pipelines, AES and NS stages are typically cascaded after linear AEC, resulting in a modular architecture where echo cancellation, residual echo suppression, and noise suppression are treated as separate processing tasks.

Despite their effectiveness, these approaches remain dependent on statistical assumptions regarding speech and interference characteristics, motivating the development of hybrid and

deep learning-based enhancement systems capable of jointly modeling more complex acoustic environments.

3.1.3 Limitations of Classical Methods

A major limitation arises from the assumption of a predominantly linear acoustic propagation model, where the echo path is represented using a linear time-invariant system. In real-world communication devices, however, non-linear distortions introduced by loudspeakers, microphones, power amplifiers, and clipping effects can significantly violate this assumption, resulting in residual echo components that are difficult to model using conventional adaptive filtering techniques [1].

Another important challenge in classical AEC systems is sensitivity to double-talk conditions, where both near-end and far-end speakers are simultaneously active. In such scenarios, adaptive filters may diverge if they continue updating using error signals that are no longer representative of the echo path.

To address this issue, Double-Talk Detection (DTD) mechanisms are commonly employed to identify periods of near-end speech activity and regulate the adaptation process of the adaptive filter. A widely studied approach is based on coherence analysis between the microphone signal and the far-end reference signal, where a decrease in inter-signal coherence is used as an indicator of double-talk conditions [20]. When double-talk is detected, the adaptation of the echo cancellation filter is typically frozen or slowed to maintain stability.

While this strategy improves robustness and prevents filter divergence, it introduces a fundamental trade-off. During detected double-talk periods, the adaptive filter is prevented from updating, which limits its ability to track changes in the acoustic echo path. As a result, when acoustic conditions change rapidly, this can lead to degraded echo cancellation performance immediately after adaptation is resumed.

Furthermore, changes in the acoustic environment, such as speaker movement or varying room characteristics, can alter the echo path over time and require continuous adaptation.

However, in practical systems, reliable adaptation is difficult to maintain during frequent double-talk or rapidly varying conditions, as the need to avoid divergence often forces conservative update strategies that slow convergence and reduce overall tracking accuracy [1].

Classical AES and NS techniques are similarly constrained by their reliance on statistical assumptions regarding stationarity and independence between speech and noise signals. In practical environments, background noise is often highly non-stationary and temporally varying, making accurate noise estimation challenging and potentially unstable. As a result, classical noise estimation methods may degrade in performance, leading to perceptual artifacts such as musical noise and excessive speech attenuation [31].

An additional limitation of traditional enhancement pipelines arises from their staged processing structure. In conventional systems, AEC, AES, and NS are treated as separate processing blocks operating sequentially [1]. While this modular decomposition simplifies system design, it also allows errors generated in earlier stages to propagate into subsequent processing steps. Residual echo artifacts produced by imperfect adaptive filtering may therefore negatively affect later suppression stages, while aggressive noise suppression can distort speech components that are already degraded by previous processing blocks.

Moreover, many classical spectral-domain approaches operate primarily on magnitude information while largely neglecting phase estimation [1]. Although magnitude enhancement can improve signal quality to some extent, inaccurate or unmodified phase information can limit perceptual speech quality and speech intelligibility, particularly in highly corrupted acoustic conditions.

3.2 Deep Learning for Speech Enhancement and Echo Suppression

The limitations of classical AEC, AES, and NS methods motivated a transition toward data-driven speech enhancement approaches based on deep learning. Rather than relying on

explicit linear system models or handcrafted statistical assumptions, these methods aim to learn direct mappings between corrupted microphone signals and enhanced speech representations using large datasets [48, 49].

Early deep learning-based speech enhancement methods primarily focused on estimating time-frequency masks capable of attenuating noise, interference, and residual echo components in the spectral domain. Various masking targets, including the Ideal Binary Mask (IBM), Ideal Ratio Mask (IRM), and Phase-Sensitive Mask (PSM), were later explored to improve speech reconstruction quality and perceptual performance [50].

As research in deep learning-based enhancement progressed, neural architectures evolved from shallow feedforward networks toward increasingly sophisticated structures capable of modeling both spectral and temporal dependencies in speech signals. CNN enabled improved local time-frequency feature extraction, while recurrent architectures such as RNNs and LSTM networks provided enhanced temporal modeling for non-stationary acoustic conditions [9, 10]. More recent systems further integrated encoder-decoder frameworks, skip connections, and frequency-time recurrent modeling strategies to improve speech reconstruction performance under challenging echo and noise environments [10, 15].

In parallel with architectural developments, modern enhancement systems progressively transitioned from magnitude-only processing toward complex-domain formulations that jointly estimate both magnitude and phase information. Complex spectral masking and mapping approaches were shown to improve perceptual speech quality and reconstruction accuracy, particularly in highly reverberant and nonlinear acoustic scenarios where phase distortion becomes increasingly significant [21].

The following sections review the major methodological developments in deep learning-based speech enhancement and acoustic echo suppression, beginning with spectral masking techniques and progressing toward complex-domain processing, temporally-aware neural architectures, and modern hybrid AEC systems.

3.2.1 Spectral Masking Approaches

One of the earliest and most widely adopted deep learning paradigms for speech enhancement and AES is spectral masking. In these approaches, neural networks are trained to estimate a time-frequency mask that selectively attenuates undesired signal components while preserving the underlying speech signal [50]. Spectral masking methods became particularly attractive due to their relatively low computational complexity and their ability to operate directly in the STFT domain.

Early masking approaches commonly employed the IBM, where each time-frequency bin is classified as either speech-dominant or interference-dominant based on local signal-to-noise conditions [51]. The IBM is typically defined as

$$M_{\mathrm{IBM}}(t,f) = \begin{cases} 1, & \text{if } \mathrm{SNR}(t,f) > \theta \\ 0, & \text{otherwise} \end{cases}. \tag{3.1}$$

where θ denotes a predefined threshold and (t, f) represent the time-frequency coordinates. In this formulation, bins dominated by speech are preserved while bins dominated by interference are attenuated. Although effective for suppressing unwanted components, binary masking often introduced perceptual artifacts and musical noise due to abrupt spectral transitions.

To address these limitations, subsequent work introduced the IRM, which replaces binary decisions with a soft-valued attenuation function [52]. The IRM can be defined either directly in terms of signal and noise power, or equivalently using the local time-frequency SNR. In its power-domain form, it is expressed as:

$$M_{\mathrm{IRM}}(t,f) = \left(\frac{|S(t,f)|^2}{|S(t,f)|^2 + |N(t,f)|^2} \right)^{\beta}. \tag{3.2}$$

where $S(t, f)$ and $N(t, f)$ denote the clean speech and interference spectra, respectively. The equivalent SNR-based formulation is given by:

$$M_{\mathrm{IRM}}(t,f) = \left(\frac{\mathrm{SNR}(t,f)}{\mathrm{SNR}(t,f)+1}\right)^{\beta}, \quad \mathrm{SNR}(t,f) = \frac{|S(t,f)|^2}{|N(t,f)|^2}. \tag{3.3}$$

Here, β is a compression parameter typically chosen in the range $[0, 1]$. Unlike binary masking approaches, ratio masks apply continuous attenuation across time–frequency bins, resulting in smoother spectral estimates and improved perceptual quality of the reconstructed speech.

Further developments explored alternative masking targets, including the PSM, which incorporates phase information into the mask estimation process to improve speech reconstruction accuracy [50]. These studies demonstrated that the choice of masking target significantly influences enhancement quality, perceptual performance, and optimization behavior in supervised speech enhancement systems.

Spectral masking approaches demonstrated substantial improvements over traditional statistical enhancement techniques, particularly under non-stationary acoustic conditions where handcrafted signal models often become unreliable [50]. However, many early deep learning methods continued to operate primarily on magnitude spectral representations while relying on the noisy phase of the corrupted microphone signal during reconstruction. This limitation motivated the development of complex-domain enhancement methods capable of jointly estimating both magnitude and phase information.

3.2.2 Complex-Domain Speech Enhancement Methods

Although spectral masking approaches demonstrated significant improvements over traditional enhancement methods, many early systems operated primarily on magnitude spectral representations while reusing the phase information of the corrupted microphone signal during reconstruction. Since phase distortions can significantly affect perceptual speech quality and intelligibility, subsequent research explored complex-domain enhancement methods capable of jointly modeling magnitude and phase information.

One of the most widely adopted approaches in this domain is the CRM, originally intro-

duced for supervised speech separation and enhancement tasks in [21]. Unlike magnitude-only masking approaches, the CRM jointly modifies both the real and imaginary components of the observed spectrum, enabling simultaneous enhancement of spectral magnitude and phase information under highly corrupted acoustic conditions.

The enhanced spectrum can be expressed as

$$\hat{Y}(f,t) = M(f,t) \cdot D(f,t). \tag{3.4}$$

where $M(f,t)$ denotes the estimated complex-valued mask and $D(f,t)$ represents the microphone spectrum in the STFT domain.

The complex mask itself is typically represented as

$$M(f,t) = M_r(f,t) + jM_i(f,t). \tag{3.5}$$

where $M_r(f,t)$ and $M_i(f,t)$ correspond to the real and imaginary components of the estimated mask respectively. By jointly estimating these components, complex-domain enhancement methods can simultaneously correct both magnitude and phase distortions present in the corrupted microphone signal.

In addition to CRM-based masking, related complex spectral mapping approaches directly estimate the real and imaginary components of the clean speech spectrum rather than predicting a multiplicative mask [53]. These approaches similarly aim to improve speech reconstruction quality through phase-aware enhancement strategies.

Complex-domain formulations demonstrated improved perceptual quality and speech intelligibility compared to magnitude-only enhancement systems, particularly in highly noisy, reverberant, and nonlinear acoustic environments [21, 7]. More recent deep complex architectures, such as the Deep Complex Convolution Recurrent Network (DCCRN), further demonstrated the effectiveness of combining complex-domain processing with convolutional and recurrent neural structures for real-time speech enhancement tasks [7].

Despite these improvements, many early deep learning enhancement systems remained limited in their ability to effectively capture long-term temporal dependencies and evolving acoustic structures over time. This motivated the development of increasingly sophisticated neural architectures capable of jointly modeling temporal and spectral information.

3.2.3 Architectural Evolution of Deep Learning-Based Enhancement Systems

Convolutional and Encoder-Decoder Architectures

Early deep learning-based speech enhancement systems primarily relied on fully connected neural networks operating on frame-level spectral representations [54]. Although these approaches demonstrated improvements over traditional statistical enhancement techniques, their ability to capture localized spectral structure and contextual information remained limited due to the absence of explicit spatial modeling within the time-frequency domain.

To address these limitations, CNN-based architectures were introduced for speech enhancement and acoustic echo suppression tasks [55]. By applying convolutional processing directly to spectral representations, CNNs enabled the extraction of localized time-frequency features while significantly reducing the number of trainable parameters compared to fully connected architectures. These properties improved robustness under non-stationary acoustic conditions and allowed enhancement systems to better model local spectral correlations in speech signals.

As research progressed, encoder-decoder architectures emerged as a dominant framework for deep learning-based speech enhancement systems. Rather than operating at a single representation scale, encoder-decoder models progressively compress corrupted spectral representations into higher-level latent feature spaces before reconstructing enhanced speech signals through successive decoding stages [10]. This hierarchical processing strategy enables the network to learn increasingly abstract spectral representations while preserving

global contextual information relevant to enhancement tasks.

A major development within this class of models was the introduction of U-Net-based architectures incorporating skip connections between corresponding encoder and decoder layers [56]. These skip connections preserve fine-grained spectral information that may otherwise be lost during feature compression, improving reconstruction accuracy and perceptual speech quality. Due to their ability to jointly model local and global spectral structure, U-Net-based architectures became widely adopted in speech enhancement, dereverberation, source separation, and acoustic echo suppression systems.

Subsequent work further extended encoder-decoder architectures into the complex spectral domain. In particular, the Deep Complex U-Net proposed by Hyeong-Seok Choi et al. [57] incorporated complex-valued convolutional operations within a U-Net encoder-decoder framework, allowing the network to jointly estimate magnitude and phase components for improved speech enhancement performance. These developments demonstrated the effectiveness of encoder-decoder architectures as a flexible foundation for modern speech enhancement systems.

Despite their strong spectral modeling capability, convolutional and encoder-decoder architectures remain fundamentally limited in their ability to explicitly model long-term temporal dependencies in speech signals. This limitation becomes particularly significant in reverberant environments and AEC scenarios where acoustic structure evolves over extended temporal contexts. This lead to later research increasingly incorporated recurrent modeling strategies to better capture sequential acoustic dynamics.

Recurrent and CRN-Based Architectures

While convolutional and encoder-decoder architectures demonstrated strong capability in modeling local spectral structure, they remain limited in their ability to explicitly capture long-term temporal dependencies in speech signals. This limitation is particularly critical in AEC and speech enhancement scenarios, where speech and echo components exhibit strong

temporal correlations over extended time contexts.

To address this issue, RNN architectures were introduced into speech enhancement systems to model sequential dependencies in acoustic signals. Early work demonstrated that LSTM networks are particularly effective in mitigating vanishing gradient issues and preserving long-range temporal information, leading to improved modeling of speech continuity and non-stationary acoustic conditions [9].

Building on these ideas, Convolutional Recurrent Network (CRN) architectures were proposed to combine the strengths of convolutional feature extraction and recurrent temporal modeling within a unified framework [10]. In these systems, convolutional encoder-decoder structures are used to extract hierarchical spectral representations, while recurrent layers are applied to the latent feature space to capture temporal dependencies across successive frames. This design enables effective modeling of both local spectral structure and long-term temporal dynamics.

More recent extensions of CRN architectures, such as the DCCRN, further integrated complex-domain spectral processing with recurrent temporal modeling [7]. By operating directly on complex-valued spectral representations, these models improve phase-aware speech enhancement performance while preserving the temporal modeling advantages of recurrent structures.

Overall, recurrent and CRN-based architectures represent a key transition point in deep learning-based speech enhancement, bridging purely convolutional approaches and more advanced frequency-time and hybrid acoustic modeling frameworks.

Frequency-Time Recurrent Architectures

While convolutional recurrent architectures such as CRN provide effective joint modeling of spectral and temporal dependencies, their recurrent modeling is typically applied along a single temporal dimension. This limits their ability to explicitly capture structured dependencies across both time and frequency axes in complex acoustic environments, where

residual echo and reverberation exhibit correlated patterns across multiple spectral bands and time scales [10].

To address this limitation, frequency-time recurrent architectures such as F-T-LSTM networks were introduced for speech enhancement and AES tasks [42]. Unlike conventional recurrent approaches that primarily model sequential temporal dependencies, F-T-LSTM architectures extend recurrent modeling into both temporal and frequency dimensions, enabling more structured representation learning over time-frequency features.

In practice, these models incorporate recurrent processing along the temporal axis while also introducing frequency-aware recurrence mechanisms that allow inter-frequency dependencies to be captured more effectively. This enables improved modeling of acoustic phenomena such as residual echo, reverberation, and non-stationary interference, which often exhibit correlated structure across both time and frequency domains.

F-T-LSTM-based systems have demonstrated strong performance in residual echo suppression and speech enhancement tasks, particularly in scenarios involving highly dynamic acoustic conditions and time-varying echo paths [15]. By extending beyond purely temporal recurrence, these architectures provide a more expressive framework for time-frequency modeling compared to standard CRN-based systems.

These developments build directly upon CRN-based approaches [10], extending their ability to model structured acoustic dependencies and motivating further exploration of hybrid architectures that combine efficient adaptive filtering with deep frequency-time neural modeling.

3.3 Research Gap and Proposed Contribution

Recent deep learning-based AEC systems have demonstrated strong performance improvements over traditional signal processing pipelines by combining neural post-filtering with classical linear adaptive echo cancellation front-ends. In particular, convolutional recurrent

architectures such as CRN-based systems and their extensions have been widely adopted due to their ability to jointly model spectral and temporal dependencies in residual echo and noise signals [10, 15].

Despite these advances, three key sources of inconsistency remain across existing hybrid AEC frameworks, which limit both fair comparison and deployment relevance.

A first source of inconsistency arises from the choice of linear adaptive echo cancellation front-end. Existing studies commonly adopt different formulations such as LMS-based, RLS-based, or frequency-domain partitioned adaptive filters such as MDF-based approaches [15]. While these methods are individually well-established, they are often selected independently across different deep learning pipelines, leading to variations in system behavior that are not attributable to the neural enhancement stage itself.

A second limitation concerns the alignment between experimental setups and real-world deployment conditions. Many hybrid AEC systems are evaluated using simplified or simulation-oriented adaptive filtering configurations, which may not reflect the constraints of production-grade echo cancellation systems, such as strict latency requirements and hardware-optimized implementations. In contrast, real-world AEC deployments typically rely on highly optimized proprietary and/or partitioned filtering structures that differ from those used in many benchmark studies [5]. This mismatch reduces the extent to which reported improvements can be directly translated to practical operating conditions.

A third source of inconsistency lies in the formulation of the neural residual echo suppression stage. Many existing approaches directly estimate enhanced speech signals or residual echo components, effectively coupling echo suppression with signal reconstruction. The framework proposed in [15], for example, achieves strong performance on benchmark datasets by leveraging near-end speech estimation objectives tailored for ICASSP challenge conditions. However, such direct estimation approaches can be sensitive to acoustic mismatch and may not generalize well outside the training distribution. In contrast, alternative formulations such as CRM based methods introduce an intermediate representation in the

complex spectral domain, which decouples enhancement from direct signal reconstruction and improves robustness under non-stationary conditions [7].

These observations suggest that reported performance improvements across existing systems may be influenced by differences in linear front-end configuration, evaluation setup, and neural target formulation, making it difficult to isolate the true contribution of the enhancement model.

Motivated by these observations, this work addresses inconsistencies in both system configuration and evaluation methodology across hybrid AEC studies. In particular, a controlled experimental framework is adopted in which both the proposed deep learning system and a traditional signal processing baseline operate under an identical linear adaptive echo cancellation stage. This ensures that performance comparisons are not confounded by differences in front-end adaptive filtering implementations, allowing improvements to be attributed directly to the neural enhancement component.

The proposed neural enhancement model is built upon the convolutional recurrent architecture introduced in [15], which serves as a baseline framework for hybrid AEC systems combining convolutional feature extraction with recurrent temporal modeling. This architecture is adopted due to its strong performance in RES tasks and its suitability for real-time processing.

To adapt this baseline architecture for deployment-oriented operation, several modifications are introduced in this thesis. First, the linear adaptive echo cancellation front-end is replaced with a production-oriented AEC system based on a proprietary implementation of a PBFDAF-based architecture originally introduced in [46]. This design choice reflects realistic acoustic operating conditions and the strict latency and computational constraints of real-time embedded AEC systems, where low-RTF operation and numerical stability are critical requirements that are often not explicitly addressed in simulation-oriented frameworks. Second, the neural enhancement stage is reformulated using a CRM-based representation instead of direct signal or residual estimation. By operating on a masking objective rather

than reconstructing the near-end speech signal directly, the network is required to learn a more constrained enhancement task, which is expected to improve robustness under acoustic mismatch and varying deployment conditions. The adoption of a CRM-based formulation also enables several architectural simplifications relative to the original framework proposed in [15]. Specifically, the dual decoder structure used for separate real and imaginary signal estimation is replaced with a shared decoder that jointly predicts the real and imaginary components of the complex mask. In addition, the auxiliary VAD branch is removed, and the decoder structure is simplified through the elimination of the final skip connection and its corresponding decoder stage. These modifications are intended to better align the architecture with the adopted masking objective while reducing model complexity and improving suitability for real-time deployment. The following section details the traditional and deep learning-based systems used in this study for comparative evaluation.

Chapter 4

Proposed Hybrid AEC System

The proposed hybrid AEC system combines a classical linear adaptive filtering stage with a deep learning-based post-filter. The system first leverages the linear AEC's ability to estimate the echo path and remove the echo component using an adaptive filter based on this estimate. The resulting signals are then used as time–frequency representations for the neural network. The neural network stage is subsequently used to model components that are not well captured by linear AEC, such as residual distortions, noise, nonlinearities, and double-talk conditions. The model then produces a CRM mask, which is applied to the microphone signal to obtain a near-end speech estimate. The overall architecture, including the linear AEC stage, the deep learning post-filter, and the loss function used in training, are described in detail in the following. This design is influenced by multi-task deep residual echo suppression frameworks with echo-aware loss [15].

4.1 System Overview

The overall system is illustrated in Figure 4.1, and the following description walks through the processing pipeline from left to right.

The first stage of the AEC system operates on the microphone signal $d[n]$ and the far-end reference signal $x[n]$. For effective adaptive filtering, these signals are temporally aligned

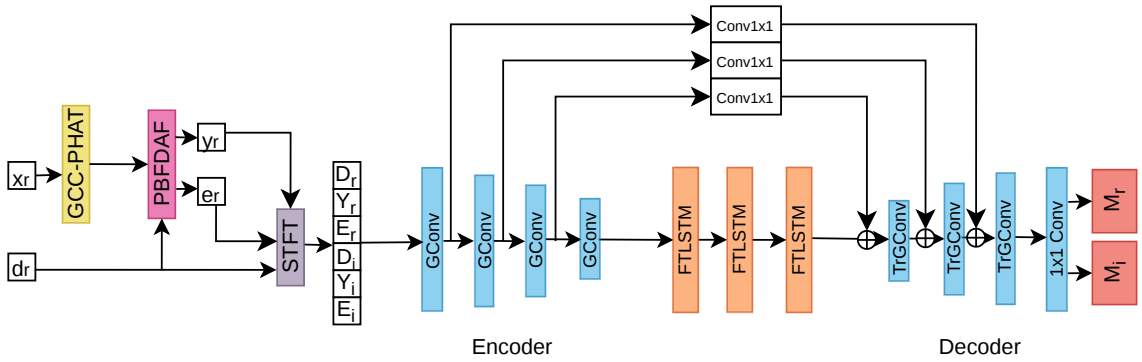


Figure 4.1: Overall hybrid AEC system architecture. The pipeline consists of a GCC-PHAT alignment stage, a PBFDAF, and a deep learning post-filter operating in the time-frequency domain. The neural model consists of a convolutional encoder, F-T-LSTM bottleneck, and decoder producing a CRM.

using a GCC-PHAT-based delay estimation stage to compensate for propagation delays introduced by the acoustic environment [58]. This preprocessing improves convergence stability by ensuring proper temporal alignment.

Following alignment, linear echo cancellation is performed using a proprietary implementation of a PBFDAF-based AEC framework, provided by an industry partner. Partitioned frequency-domain adaptive filters are widely used in real-time systems due to their efficiency in modeling long acoustic impulse responses [59]. The filter produces an echo estimate $y[n]$, which is subtracted from the microphone signal to yield the residual signal:

$$e[n] = d[n] - y[n].$$

where $e[n]$ contains near-end speech, residual echo, and background noise.

Although the linear AEC reduces dominant echo components, residual artifacts remain due to nonlinearities, time-varying acoustic paths, and double-talk conditions. Therefore, a deep learning-based post-filter is applied for further suppression of residual echo and noise.

The signals $d[n]$, $y[n]$, and $e[n]$ are transformed into the time-frequency domain using the STFT [19]. These complex representations are stacked into feature tensors containing real and imaginary components, enabling joint magnitude and phase modeling.

The deep learning post-filter consists of a convolutional encoder, a F-T-LSTM bottleneck, and a transposed convolutional decoder. This architecture captures both local spectral structure and long-range temporal dependencies [42].

The network outputs a complex ratio mask (CRM):

$$M[k,m] = M_r[k,m] + jM_i[k,m].$$

which is applied to the microphone spectrum:

$$\hat{S}[k,m] = M[k,m] \cdot D[k,m].$$

Finally, the enhanced speech signal is reconstructed using inverse STFT:

$$\hat{s}(n) = \text{ISTFT}(\hat{S}[k,m]).$$

4.2 Linear AEC Stage

4.2.1 PBFDAF Linear Acoustic Echo Cancellation

The implementation used in this work is a proprietary PBFDAF-based AEC framework inspired by the approach presented in [46] and provided by an industry partner. Therefore, specific implementation details and optimization strategies are not covered. Nevertheless, the overall processing structure follows the general principles employed in partitioned frequency-domain adaptive echo cancellation systems.

In AEC, the echo signal is typically modeled using an adaptive FIR filter, $\hat{h}[n]$, where the estimated echo is obtained by convolving the far-end reference signal $x[n]$ with the adaptive filter:

$$\hat{y}[n] = \sum_{i=0}^{M-1} \hat{h}[i]x[n-i]. \tag{4.1}$$

The adaptive filter must model the loudspeaker-room-microphone impulse response, which can span thousands of taps in reverberant environments such as vehicle cabins, conference rooms, and hands-free communication systems [46]. As the filter length M increases, direct time-domain convolution and adaptive coefficient updates become increasingly computationally expensive, making real-time implementation challenging on embedded hardware platforms.

To reduce computational complexity, filtering can instead be performed in the frequency domain using the FFT. By transforming both the input signal and adaptive filter into the frequency domain, convolution is replaced with element-wise multiplication according to the convolution theorem. This reduces computational complexity from approximately $\mathcal{O}(M^2)$ for direct convolution to approximately $\mathcal{O}(M \log M)$ when implemented using FFT-based processing [2, 19]. Frequency-domain adaptive filtering is therefore particularly attractive for long adaptive filters encountered in AEC applications.

A straightforward frequency-domain adaptive filter, however, requires processing large blocks of data to model long impulse responses, introducing additional latency that may be unsuitable for real-time interactive communication systems [2]. To address this limitation, the PBFDAF partitions the adaptive filter into multiple smaller sub-filters that are processed independently in the frequency domain. This partitioned structure maintains the computational advantages of frequency-domain processing while reducing latency and enabling efficient implementation of long adaptive filters.

The algorithm employs block-based processing using an overlap-save style framework [46]. The far-end reference and microphone signals are segmented into overlapping blocks and transformed into the frequency domain, where filtering and adaptation are performed on a block-by-block basis. The partitioned structure enables efficient fast operations while maintaining suitability for real-time operation.

The approach presented in [46] further improves computational efficiency by limiting the adaptive filter length rather than attempting to model the entire acoustic impulse response

using a very large number of coefficients. Instead, the adaptive filter is designed to capture the dominant early echo components that contain most of the acoustic energy. Remaining late residual echo components are subsequently attenuated by additional post-processing stages within the overall hands-free communication system.

The primary motivation for selecting this class of algorithm is its scalability and computational efficiency for long echo paths commonly encountered in practical acoustic environments. Consequently, PBFDAF-based linear echo cancellation remains a widely adopted front-end solution for real-time hands-free communication systems.

4.2.2 Partitioning of the Adaptive Filter

As discussed previously, the PBFDAF improves computational efficiency by dividing the adaptive filter into multiple smaller partitions that are processed independently in the frequency domain [46]. Rather than adapting a single long filter, the overall impulse response is represented by a collection of shorter sub-filters.

Assuming an overall adaptive filter length of M, the filter is divided into L partitions of length N, such that

$$M = L \times N. \tag{4.2}$$

Each partition models a specific temporal region of the acoustic echo path. During processing, the contribution of every partition is computed independently in the frequency domain and combined to form the final echo estimate.

This structure allows the processing delay to be determined primarily by the partition length N rather than the overall filter length M, while also distributing the computational load across multiple smaller transforms. This allows this approach to remain scalable for the long impulse responses encountered in practical acoustic environments [59].

The approach presented in [46] further improves computational efficiency by limiting the

modeled impulse response length rather than attempting to estimate the full acoustic path.

An additional advantage of the partitioned framework is that different sections of the impulse response can be adapted independently. As implemented in [46], separate adaptation parameters may be applied to individual partitions, improving tracking performance under time-varying acoustic conditions and during double-talk periods.

Mathematical Decomposition

Rather than computing the output of a single long adaptive filter, the estimated echo signal is expressed as the sum of the contributions from all partitions:

$$\hat{y}[n] = \sum_{l=0}^{L-1} \hat{y}_l[n]. \tag{4.3}$$

where the output of the l-th partition is given by

$$\hat{y}_l[n] = \sum_{i=0}^{N-1} \hat{h}_l[i]\, x[n - lN - i]. \tag{4.4}$$

Here, $\hat{h}_l[i]$ denotes the coefficient vector associated with the l-th partition of the adaptive filter. The term lN introduces the appropriate delay so that each partition models a distinct temporal region of the acoustic impulse response.

This decomposition allows a long acoustic echo path to be represented as a structured collection of shorter adaptive sub-filters. In the PBFDAF, each partition is transformed into the frequency domain and processed using block-based convolution [1]. The resulting partition outputs are accumulated to form the final estimated echo signal.

Interpretation as Block Convolution

From a signal processing perspective, the partitioning scheme converts a single long convolution into a sum of multiple shorter convolutions. Instead of applying one full-length adaptive filter directly in the time domain, the PBFDAF decomposes the filtering operation

into several localized convolutions corresponding to the individual filter partitions [46].

Each partition operates on a delayed version of the input signal and produces a partial echo estimate associated with a specific temporal region of the acoustic path. The outputs of all partitions are then combined to form the final estimated echo signal. This process is equivalent to reconstructing the response of the original long adaptive filter using multiple shorter sub-filters processed independently in the frequency domain.

To preserve the temporal structure of the original system, the output of each partition is aligned according to its associated delay lN before summation. The resulting output therefore remains mathematically equivalent to the linear convolution produced by a single full-length adaptive filter, while benefiting from the reduced computational complexity of block frequency-domain processing.

This restructuring is fundamental to the efficiency of the PBFDAF. By dividing the convolution into smaller frequency-domain operations, the algorithm distributes the computational workload across multiple smaller transforms instead of relying on a single large transform [59]. This significantly reduces computational peaks, improves scalability for long impulse responses, and enables practical real-time execution on embedded Digital Signal Processing (DSP) hardware.

4.2.3 Block-Based Input Representation

In block-based adaptive filtering implementations, both the input and microphone signals are segmented into overlapping frames prior to transformation into the frequency domain. This enables efficient convolution through the use of the FFT, replacing time-domain convolution with frequency-domain multiplication [2, 1].

Since each filter partition $\hat{h}_l(n)$ represents a delayed portion of the overall acoustic impulse response, the adaptive filtering process relies on maintaining a history of previous input blocks. This allows each partition to operate on the appropriately time-aligned spectral representation of the input signal within the block-based framework.

To ensure equivalence with linear convolution and avoid circular convolution, each input block of length N is typically zero-padded prior to transformation into the frequency domain. The block is defined as:

$$x_k(n), \quad n = 0, 1, \ldots, N-1. \tag{4.5}$$

After zero-padding, the frequency-domain representation is obtained using the FFT:

$$X_k = \mathrm{FFT}\{x_k(n)\}. \tag{4.6}$$

This formulation follows the standard overlap-save processing framework commonly used in frequency-domain adaptive filtering [19, 1].

In the partitioned structure, the estimated echo spectrum at block index k is given by the sum of contributions from all filter partitions:

$$Y_k = \sum_{l=0}^{L-1} \hat{H}_l(k) \cdot X_{k-l}. \tag{4.7}$$

where:

- $\hat{H}_l(k)$ denotes the frequency-domain adaptive filter coefficients for the l-th partition,
- X_{k-l} represents the delayed input spectrum associated with that partition.

This partitioned convolution structure is consistent with frequency-domain adaptive filtering frameworks such as MDF and related PBFDAF implementations [22, 23].

A key advantage of this architecture is that the processing latency is primarily determined by the block length N rather than the total filter length M. This enables long acoustic impulse responses to be modeled efficiently while maintaining real-time constraints in hands-free communication systems [46].

Within this general framework, the estimated echo signal is computed as the sum of contributions from multiple filter partitions operating on delayed spectral representations

of the input signal. Each partition operates on a different segment of the acoustic impulse response, enabling efficient modeling of long reverberation tails without requiring a single long time-domain filter.

4.2.4 Frequency-Domain Filtering

Once the frequency-domain input buffer is populated and the filter partitions $\hat{H}_l(k)$ are initialized, the estimated echo spectrum Y_k is computed as a weighted summation of the buffered input spectra:

$$Y_k = \sum_{l=0}^{L-1} X_{k-l} \cdot \hat{H}_l(k). \tag{4.8}$$

where X_{k-l} denotes delayed input spectra stored in a circular buffer, and $\hat{H}_l(k)$ represents the frequency-domain coefficients associated with the l-th adaptive filter partition. This formulation corresponds to the standard partitioned convolution structure used in frequency-domain adaptive filtering frameworks such as MDF and related PBFDAF implementations.

After the spectral-domain filtering operation, the estimated echo spectrum is transformed back into the time domain using the inverse FFT:

$$\hat{y}_k = \mathrm{IFFT}(Y_k). \tag{4.9}$$

Because multiplication in the frequency domain corresponds to circular convolution for non periodic signals, overlap-save processing is typically used to obtain the equivalent linear convolution result. In this work, a block length of $2N$ is used within the Overlap-Save (OLS) framework [2]. After the inverse transform, only the last N samples of the output block are retained, while the first N samples are discarded, as they contain wrap-around artifacts introduced by circular convolution.

This overlap-save structure ensures that the block-based frequency-domain implementation remains equivalent to linear convolution while maintaining computational efficiency for long adaptive filters.

4.2.5 Error Signal Computation and Adaptive Filter Update

Once the estimated echo signal $\hat{y}_k$ is reconstructed in the time domain, the residual error signal is computed as:

$$e_k = d_k - \hat{y}_k. \tag{4.10}$$

where d_k denotes the microphone signal block containing near-end speech, background noise, and acoustic echo components. The residual signal therefore represents the portion of the microphone signal not explained by the current adaptive filter estimate.

For frequency-domain adaptive filtering, the residual error signal is transformed back into the spectral domain using a zero-padded FFT representation:

$$E_k = \mathrm{FFT}\left(\begin{bmatrix}\mathbf{0}_N \\ e_k\end{bmatrix}\right). \tag{4.11}$$

This step is consistent with standard overlap-save-based adaptive filtering implementations, where zero-padding ensures proper alignment between the error spectrum and the partitioned input spectra during coefficient updates [2, 1].

The adaptive filter partitions are then updated using a frequency-domain LMS-type adaptation rule with partition-dependent step-size control:

$$\hat{H}_l^{(k+1)} = \hat{H}_l^{(k)} + \mu_l^{(k)} \cdot X_{k-l}^* \odot E_k. \tag{4.12}$$

where:

- $\mu_l^{(k)}$ is the partition-dependent adaptation gain controlling convergence speed and stability,
- X_{k-l}^* denotes the complex conjugate of the delayed input spectrum associated with partition l,
- $\odot$ denotes element-wise multiplication,

- E_k is the frequency-domain representation of the residual error signal.

This update structure corresponds to the general class of frequency-domain LMS-type adaptation used in partitioned block adaptive filtering frameworks such as MDF and related PBFDAF implementations [22, 23].

In the system proposed by Estermann and Kaelin [46], partitioned adaptation is used at an architectural level to improve robustness in real-time hands-free echo cancellation systems. Rather than enforcing a single uniform adaptation gain across all partitions, the framework allows adaptation to vary across segments of the impulse response depending on signal conditions such as double-talk and transient interference. This enables more stable tracking of the acoustic path in practical environments.

In practice, partition-dependent step-size control can also be interpreted as assigning different adaptation sensitivities to different regions of the acoustic impulse response. Partitions associated with early, high-energy reflections typically benefit from stronger adaptation, while later partitions corresponding to weaker, noise-sensitive components may require more conservative updates. This improves numerical stability and reduces the risk of divergence under non-stationary acoustic conditions.

4.2.6 Practical Justification Compared to MDF and RLS

While multiple adaptive filtering approaches exist for AEC, including MDF and RLS-based methods such as those used in the hybrid AEC framework presented in [15] that inspired this work, the partitioned frequency-domain architecture adopted in this work provides a more practical balance between computational efficiency, convergence stability, latency, and implementation complexity for real-time embedded deployment.

The hybrid architecture proposed in [15] combines classical adaptive filtering with deep learning-based residual suppression; however, the linear echo cancellation stage primarily relies on conventional Weighted Recursive Least Squares (wRLS) and textbook MDF-style adaptive filtering structures. Although these approaches can provide strong echo cancellation

performance under controlled conditions, practical deployment in embedded automotive systems introduces additional constraints related to computational determinism, memory usage, numerical stability, and low-latency real-time execution.

First, considering wRLS, the algorithm provides very rapid convergence for highly correlated input signals such as speech. However, its computational and memory requirements scale quadratically with filter length, making it impractical for the long adaptive filters typically required in real-time acoustic echo cancellation systems, particularly on resource-constrained embedded platforms [14]. In addition, wRLS implementations are generally more sensitive to numerical instability and finite-precision effects, particularly in fixed-point or resource-constrained hardware environments.

When looking at MDF while it significantly reduces computational complexity compared to purely time-domain adaptive filtering, several practical limitations arise in real-time embedded implementations. Since frequency-domain multiplication inherently produces circular convolution, overlap-save constraints must be enforced to preserve equivalence with linear convolution. Therefore, the coefficient update stage requires repeated FFT/Inverse Fast Fourier Transform (IFFT) operations across all filter partitions to remove time-domain aliasing effects [23]. For long acoustic impulse responses, the increasing number of filter partitions can substantially increase memory usage, processing overhead, and execution-time variability in low-latency embedded systems [18].

In addition, textbook MDF formulations typically employ a global normalization or adaptation strategy across the entire filter structure. While effective under stationary conditions, this global adaptation behaviour can become less robust in highly dynamic acoustic environments such as automotive cabins, where double-talk, transient disturbances, and rapid acoustic path variations occur frequently. Under these conditions, portions of the adaptive filter may require different adaptation behaviour to maintain stability and tracking performance.

The partitioned frequency-domain framework proposed by Estermann and Kaelin in [46]

was specifically designed to address many of these practical real-time limitations in hands-free communication systems. Rather than indefinitely increasing the adaptive filter length to fully model the complete acoustic impulse response, the architecture intentionally utilizes a shorter adaptive filter focused on modeling the dominant early echo components. Residual late echo components are subsequently attenuated using additional suppression stages integrated within the overall system architecture. This reduces computational complexity and memory requirements while still maintaining effective perceptual echo suppression performance.

Furthermore, the framework incorporates localized adaptation control across the partitioned filter structure. Instead of relying solely on a single global adaptation constraint, individual filter partitions can independently modify or suspend adaptation depending on local signal conditions. This partition-dependent adaptation improves numerical stability and robustness during double-talk or transient interference conditions while allowing other partitions to continue tracking the acoustic path. By combining partitioned frequency-domain processing with localized adaptation control, the framework achieves improved convergence behaviour and tracking performance while maintaining substantially lower computational complexity than wRLS and MDF based methods.

For these reasons, the proposed system adopts a proprietary partitioned frequency-domain AEC framework inspired by the real-time architectural principles presented in [46], rather than relying on conventional textbook MDF or computationally intensive RLS implementations. This provides a more practical and deployment-oriented solution for low-latency automotive hands-free communication environments while maintaining compatibility with the subsequent deep learning-based residual echo and noise suppression stages.

4.2.7 Summary of the PBFDAF Processing Chain

The linear AEC stage operates on a block-by-block basis using partitioned frequency-domain adaptive filtering. The processing sequence for each iteration k is summarized as follows:

1. **Input Buffering:** Construct the overlap-save buffer using the current far-end input

block x_k.

2. **Spectral Transformation:** Apply the $2N$-point FFT to obtain the input spectrum X_k and update the spectral history buffer containing previous input blocks.

3. **Partitioned Filtering:** Compute the estimated echo spectrum:

$$Y_k = \sum_{l=0}^{L-1} X_{k-l} \cdot \hat{H}_l^{(k)}. \tag{4.13}$$

4. **Time-Domain Reconstruction:** Apply the IFFT to obtain the time-domain estimated echo signal $\hat{y}_k$.

5. **Valid Sample Extraction:** Retain the valid N samples of the reconstructed block and discard transient samples introduced by circular convolution effects.

6. **Residual Computation:** Compute the error signal:

$$e_k = d_k - \hat{y}_k. \tag{4.14}$$

7. **Error Transformation:** Transform the residual signal into the frequency domain to obtain the error spectrum E_k.

8. **Adaptive Update:** Update each filter partition using a frequency-domain stochastic gradient update:

$$\hat{H}_l^{(k+1)} = \hat{H}_l^{(k)} + \mu_l^{(k)} \cdot X_{k-l}^* \odot E_k. \tag{4.15}$$

This processing structure follows the general partitioned frequency-domain adaptive filtering framework commonly used in real-time AEC systems [1, 23, 2]. While the formulation is consistent with standard PBFDAF-style adaptive filtering theory, the system-level design adopted in this work is inspired by the real-time partitioned architecture described in [46],

which emphasizes practical constraints such as computational determinism, bounded latency, and robustness in hands-free communication scenarios.

The output of the linear AEC stage provides a decomposition of the microphone signal into an estimated echo component $\hat{y}(n)$ and a residual signal $e(n)$. These component are then used as the inputs to the deep learning post filter.

In particular, the post filter uses the time-frequency representations of $d(n)$, $\hat{y}(n)$, and $e(n)$ to learn a nonlinear mapping that refines the linear AEC output. In this hybrid framework, the PBFDAF provides a computationally efficient linear prior, while the neural network models residual nonlinear and data-dependent components that are not captured by the adaptive filter.

4.3 Deep Learning Post-Filter

The second stage of the proposed architecture consists of a deep learning-based post-filter. The primary objective of this module is to refine the outputs of the linear AEC stage by jointly suppressing residual echo, acoustic nonlinearities, and ambient background noise, while simultaneously minimizing distortion of the near-end speech components.

Compared to traditional DSP post-filtering approaches, such as Wiener filtering or spectral subtraction [30, 27], the proposed model avoids explicit statistical assumptions on signal and noise distributions. Conventional methods rely on PSD estimation and are sensitive to non-stationary interference, often introducing artifacts such as musical noise [60].

In contrast, the deep learning model learns a nonlinear mapping from multi-channel spectral inputs, enabling robust suppression of residual echo and noise even under double-talk and spectral overlap conditions. By leveraging the microphone signal, linear echo estimate, and residual error, the model performs a structured refinement of the AEC output.

The following sections describe the signal preprocessing pipeline, neural network architecture, and complex-valued masking strategy used to achieve this behavior.

4.3.1 Feature Extraction and Signal Representation

To interface the linear AEC outputs with the deep learning post-filter, the signals are pre-processed using an STFT configuration and a multi-channel stacking strategy. This provides a normalized, phase-aware representation of the microphone signal and linear AEC outputs.

Spectral Analysis and Parameters

The system operates at a sampling rate of 16 kHz, providing sufficient bandwidth for wide-band speech processing. A 20 ms Hann window ($L = 320$ samples) is used, balancing the quasi-stationarity assumption of speech with temporal resolution.

A 50% overlap is applied ($R = 160$ samples), satisfying the Constant Overlap-Add (COLA) condition for stable reconstruction. Each frame is processed using a 512-point FFT, resulting in 257 non-redundant frequency bins with a resolution of 31.25 Hz.

This configuration provides an effective trade-off between spectral resolution and computational efficiency.

For each frame m, the following frequency-domain signals are computed:

- **Microphone signal** $D[k, m]$**:** mixture of speech, echo, and noise.
- **Echo estimate** $Y[k, m]$**:** output of the linear PBFDAF.
- **Residual signal** $E[k, m]$**:** computed as $D[k, m] - Y[k, m]$, representing the echo-suppressed signal.

Input Feature Selection and Comparison

The final model uses the microphone signal, echo estimate, and residual signal as inputs. Alternative configurations replacing $Y(k, m)$ with the far-end reference $X(k, m)$ were also evaluated, following [15].

However, incorporating the raw far-end signal led to overly aggressive suppression and increased distortion of near-end speech. Without guidance from the linear AEC output,

the network must implicitly learn both echo cancellation and speech preservation, which complicates the learning objective.

In contrast, using $Y(k, m)$ and $E(k, m)$ provides a structured decomposition of the signal. The linear PBFDAF removes the dominant linear echo component, allowing the network to focus on residual artifacts such as nonlinear echo, double-talk interference, and background noise. This simplifies the task to a refinement problem.

A detailed comparison of these input configurations is provided in Section 6.

Normalization and Input Stacking

Each real ($\Re$) and imaginary ($\Im$) components are standardized to zero mean and unit variance:

$$\tilde{S}_{r,i} = \frac{S_{r,i} - \mu_{S_{r,i}}}{\sigma_{S_{r,i}} + \epsilon}. \tag{4.16}$$

where $\epsilon = 10^{-8}$ ensures numerical stability.

The resulting normalized features are stacked along the channel dimension to form the input tensor $\mathbf{X} \in R^{B\times 6\times F\times T}$, where B denotes the batch size, F denotes the number of frequency bins, and T denotes the number of time frames. During training, multiple input segments are processed simultaneously to improve computational efficiency and stabilize gradient estimation. The batch size B therefore represents the number of spectrogram samples processed together in a single forward and backward pass through the network. The frequency dimension F corresponds to the spectral bins produced by the STFT, while the temporal dimension T represents the sequence of analysis frames spanning the input signal. The resulting 6-channel representation provides the model with simultaneous access to the raw mixture, the linear AEC output, and the residual signal:

$$\mathbf{X} = \{D_r, D_i, Y_r, Y_i, E_r, E_i\}. \tag{4.17}$$

Including both the linear estimate Y and residual E allows the network to explicitly learn

the discrepancy between the linear AEC model and the true acoustic environment, enabling better modelling of nonlinear and non-stationary components not captured by the PBFDAF.

Streaming Block Processing and State Management

Unlike offline models, the proposed post-filter operates in a causal streaming manner to support real-time AEC deployment. Processing is performed in the time-frequency domain using a sliding-window strategy designed to balance temporal context, enhancement quality, and the strict latency constraints of automotive communication systems. To provide temporal context for the deep learning model while maintaining causal operation, spectral frames are grouped into chunks of $C_{size} = 16$ frames using a rolling buffer mechanism. The model operates exclusively on current and past frames without the use of future-frame lookahead, ensuring that no additional non-causal latency is introduced by the neural network. Once the buffer is initialized, newly arriving frames replace the oldest frames in the buffer, enabling continuous streaming inference. A high-overlap configuration of $O = 14$ frames was selected, resulting in a stride of 2 frames between consecutive inference windows. Since each STFT frame corresponds to 10 ms, the network performs inference updates every 20 ms while preserving a total receptive field of approximately 160 ms. Importantly, the overlap configuration affects the temporal update rate and computational cost of the model, but does not increase the algorithmic latency, which remains dominated by the causal STFT analysis window duration. While lower-overlap configurations were also evaluated to reduce computational complexity, the 16×14 sliding-window configuration was selected as the final design due to its improved temporal continuity and enhanced modeling of long-tail acoustic reflections commonly encountered in automotive cabin environments.

4.3.2 Architecture

The proposed architecture follows a convolutional encoder–decoder design with an F-T-LSTM bottleneck, enabling the model to capture both local time–frequency structure and long-range

temporal dependencies. The design is inspired by the U-Net topology [61, 15, 57], which employs a symmetric contracting and expanding path for hierarchical feature extraction. However, modifications were made to the traditional U-Net topology to improve real time performance.

Encoder

The encoder transforms the multi-channel acoustic input into a compact hierarchical representation that preserves both local spectral structure and broader contextual information, providing an efficient interface to the bottleneck for subsequent sequential modeling [61, 57].

The input to the encoder is a multi-channel feature tensor:

$$\mathbf{x} \in R^{B \times C \times F \times T}. \tag{4.18}$$

where $C = 6$ corresponds to stacked acoustic features derived from the microphone signal $d[n]$, the estimated echo $\hat{y}[n]$, and the residual signal $e[n]$, all represented in the time–frequency domain using the STFT.

The encoder progressively increases channel dimensionality while reducing resolution along both the time and frequency axes. Through a series of strided convolutional layers, the encoder performs learned spectro-temporal downsampling, allowing the receptive field to grow and enabling the extraction of increasingly abstract acoustic representations. These representations capture higher-level characteristics of the input signal, such as residual echo structure and non-stationary noise patterns, and are subsequently passed to the bottleneck for temporal modeling.

Gated Convolutional Feature Extraction

A key element of the encoder is the use of GConv blocks as the fundamental building units. These differ from standard convolutional layers by using a dual-branch structure consisting

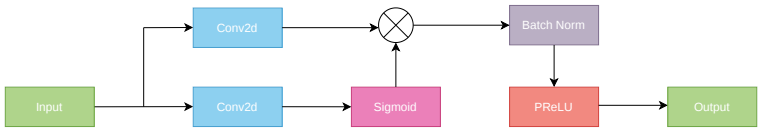


Figure 4.2: Structure of the GConv block used throughout the encoder. The feature branch extracts spectral representations while the gating branch learns a signal-dependent mask that selectively controls information flow.

of a feature extraction path and a gating path, as proposed in [62].

The internal structure of the GConv block used throughout the encoder is illustrated in Figure 4.2. The input feature map is processed by two parallel convolutional branches. The first branch generates the feature representation, while the second branch produces a gating mask that is passed through a sigmoid activation function to constrain its values between 0 and 1. The outputs of the two branches are then multiplied element-wise, allowing the network to dynamically regulate the amount of information propagated to subsequent layers. Following the gating operation, batch normalization and a Parametric Rectified Linear Unit (PReLU) activation are applied to improve training stability and representational capacity.

The gating mechanism acts as a learned, signal-dependent filter that modulates feature flow based on local time-frequency context. This is particularly useful in AEC scenarios, where residual echo and background noise often overlap spectrally with near-end speech. By learning adaptive gating masks, the network selectively emphasizes informative components while suppressing interference, improving robustness in non-stationary acoustic environments[63, 15, 64].

Mathematical Formulation

For an input feature map $\mathbf{x} \in R^{B\times C_{in}\times F\times T}$, the GConv operation computes two parallel transformations.

Feature Extraction Branch: The candidate acoustic features are extracted using a learnable convolutional kernel $\mathbf{W}_f$:

$$\mathbf{f} = \mathbf{W}_f * \mathbf{x} + \mathbf{b}_f. \tag{4.19}$$

where $*$ denotes convolution and $\mathbf{b}_f$ is the bias term.

Gating Branch: In parallel, a gating signal is generated using a second convolutional kernel $\mathbf{W}_g$ followed by a sigmoid activation:

$$\mathbf{g} = \sigma(\mathbf{W}_g * \mathbf{x} + \mathbf{b}_g). \tag{4.20}$$

The sigmoid function constrains the output to $(0, 1)$, enabling it to act as a soft, data-dependent mask over the extracted features.

Output Integration: The outputs of the feature extraction and gating branches are combined through element-wise multiplication:

$$\mathbf{z} = \mathbf{f} \odot \mathbf{g}. \tag{4.21}$$

where $\odot$ denotes element-wise multiplication. The resulting feature map is then normalized using batch normalization and passed through a PReLU activation function $\phi(\cdot)$:

$$\mathbf{y} = \phi\left(\mathrm{BN}(\mathbf{z})\right). \tag{4.22}$$

where $\mathrm{BN}(\cdot)$ denotes the batch normalization operation.

In the context of the AEC post-filter, this multiplicative interaction enables adaptive feature modulation, allowing the network to emphasize or attenuate spectro-temporal regions according to the input characteristics. This mechanism can facilitate the suppression of residual echo and other interference components while preserving speech-related features.

Design Specifications

With the core structural components defined, it is important to outline the specific design parameters of the model. These parameters are chosen to balance computational efficiency with the requirements of real-time acoustic modeling. For both the feature extraction and gating branches within the GConv blocks, a fixed kernel size of 3×3 is used with unit zero-padding to preserve spatial resolution in the early stages of the network.

Choice of Kernel Size (3×3) The use of 3×3 kernels across all convolutional layers is motivated by the structure of time-frequency acoustic signals. In an STFT representation, speech and echo exhibit strong local correlations along both frequency and time axes, which can be effectively captured using compact receptive fields.

In the frequency domain, speech is characterized by harmonic structure, where energy is concentrated around fundamental frequencies and their harmonics. The use of 3×3 kernels is motivated by the local structure of time-frequency acoustic signals. In the STFT domain, both speech and echo exhibit strong local correlations across adjacent time frames and frequency bins, which can be effectively captured using compact receptive fields.

This choice provides:

- Efficient parameterization suitable for real-time processing.
- An inductive bias toward local time-frequency structure consistent with speech and echo characteristics.

Padding Strategy ($P = 1$) A symmetric padding size of $P = 1$ is used for all 3×3 convolutions to preserve spatial dimensions. For stride $S = 1$, the output dimension is given by:

$$D_{out} = \left\lfloor \frac{D_{in} - K + 2P}{S} \right\rfloor + 1. \tag{4.23}$$

which simplifies to $D_{out} = D_{in}$ under this configuration.

This choice is commonly adopted in convolutional encoder–decoder architectures to preserve spatial resolution across layers and maintain alignment of structured feature maps [61]. It is particularly important in time–frequency representations, where preserving bin-level correspondence helps maintain temporal and spectral consistency across layers, especially for transient speech and early echo components.

Batch Normalization (BN) Following the gating operation, BN is applied to stabilize feature distributions. Due to the high dynamic range of STFT-based inputs and the multiplicative nature of gated convolutions, feature variance can increase across layers.

BN normalizes activations to reduce internal covariate shift, improving training stability and allowing more consistent gradient flow through deep layers [65].

PReLU Activation A PReLU activation function is applied after normalization:

$$f(x) = \max(0, x) + \alpha \min(0, x). \tag{4.24}$$

Unlike standard Rectified Linear Unit (ReLU), PReLU retains negative activations through a learnable parameter α. This is useful in this architecture because normalized gated features are centered around zero, and negative-valued components can still contain meaningful acoustic information. This improves flexibility in modeling low-energy speech and residual echo components while supporting stable gradient propagation [66].

Benefits of Gated Convolutions Gated convolutional blocks provide two key advantages in time-frequency speech processing:

- **Implicit masking behavior:** The gating mechanism performs adaptive feature suppression, which can be interpreted as a learned time-frequency mask that attenuates residual echo and noise while preserving speech components.

- **Robustness to non-stationarity:** The gating function allows the network to adapt to time-varying acoustic conditions, such as double-talk and changing noise floors, by dynamically modulating feature importance.

Encoder Structure

With the individual Gated Convolutional blocks defined, the next step is to describe the overall encoder architecture and its hierarchical role within the system. The encoder is composed of a stack of GConv blocks with progressively increasing channel depth and decreasing spatial resolution, as detailed in Table 4.1.

Layer	Input Ch.	Output Ch.	Stride	Padding
L1	6	32	1	1
L2	32	64	2	1
L3	64	128	2	1
L4	128	256	2	1

Table 4.1: Encoder architecture showing progressive channel expansion and time–frequency downsampling.

Initial Feature Projection The first layer maps the 6-channel input tensor to 32 feature maps using stride $S = 1$. This stage serves as an initial feature projection, transforming raw time–frequency inputs into a higher-dimensional representation while preserving full spectral resolution. In the context of AEC, maintaining this resolution is important for preserving fine acoustic details such as weak speech components and early echo reflections, which are critical for accurate residual suppression[57].

Hierarchical Downsampling and Feature Expansion From the second layer onward, strided convolutions ($S = 2$) are introduced to progressively reduce the time–frequency resolution while increasing channel dimensionality [57, 67]. This design achieves two main objectives:

- **Computational efficiency:** Reducing spatial resolution significantly lowers the number of operations in deeper layers and prepares the representation for the recurrent bottleneck.

- **Feature abstraction:** Increasing channel depth allows the network to encode progressively more complex acoustic structures, transitioning from local spectral patterns to broader representations of reverberation and noise.

Multi-Scale Feature Representation This progressive compression creates a multi-scale representation where network depth corresponds to increasing abstraction. Early layers focus on local spectral structure, while deeper layers (L3–L4) capture more global acoustic features such as residual echo envelopes. This hierarchical representation is well suited for AEC and AES tasks, since residual echo is inherently multi-scale, containing both short-term spectral artifacts and longer-term reverberation components.

Transition to Temporal Modeling The encoder output is specifically optimized for the F-T-LSTM bottleneck. By employing a stride of 2 in both time and frequency dimensions, the model reduces the temporal sequence length by a factor of 8. This compression reduces the computational burden of temporal modeling while preserving the key acoustic structures required for long-term echo tracking.

Finally, skip connections are introduced at all but the last encoder layer to preserve fine-grained information lost during downsampling. These connections provide the decoder with access to high-resolution spectral cues, improving reconstruction quality and supporting stable gradient propagation throughout the network [67].

Skip Connections and Multi-Scale Fusion

Skip connections provide the decoder with access to intermediate encoder representations, enabling effective multi-scale feature fusion during reconstruction. This architectural design

is widely used in speech separation and enhancement encoder–decoder networks to preserve fine-grained information lost during downsampling while improving reconstruction fidelity [67].

In the context of AEC, this multi-scale feature fusion is technically significant for several reasons:

- **Preservation of Spectral Detail:** Fine-grained speech components, such as high-frequency harmonics and transient onsets, can be attenuated during strided downsampling in the encoder. Skip connections recover these cues, enabling the decoder to reconstruct a mask that preserves near-end speech with higher perceptual fidelity [67].
- **Targeted Artifact Suppression:** By providing the decoder with access to higher-resolution encoder features, the model can more effectively identify and suppress residual echo artifacts that may otherwise be blurred in the compressed bottleneck representation [67].
- **Gradient Flow Optimization:** Skip connections improve gradient propagation by reducing the effective depth between the loss function and early encoder layers, mitigating vanishing gradients and accelerating convergence during training [67].

Transition to the Bottleneck Stage While the encoder extracts hierarchical time–frequency representations, the bottleneck is responsible for modeling long-range temporal dependencies that cannot be captured by purely convolutional processing. This is particularly important in AEC, where echo components may persist over extended time intervals due to room reverberation and acoustic propagation delays.

This motivates the use of a recurrent bottleneck that can integrate information across multiple time frames, enabling temporally consistent suppression of residual echo and noise.

F-T-LSTM Architecture and Design

The design of the F-T-LSTM module follows the original frequency–temporal decomposition idea proposed in [68], where sequential modeling is structured along separate feature axes to better exploit the inherent structure present in both the temporal and frequency domains.

This concept was later adapted to AES tasks in [42], where frequency- and time-domain dependencies are modeled explicitly through a decoupled Frequency Long Short-Term Memory (F-LSTM) and Temporal Long Short-Term Memory (T-LSTM) structure, enabling more effective suppression of non-stationary noise and residual echo.

The internal structure of the proposed F-T-LSTM layer is illustrated in Figure 4.3. The module first performs frequency-axis modeling, followed by temporal-axis modeling, allowing the network to jointly capture spectral and temporal dependencies in a structured manner.

This dual-path formulation aligns well with the requirements of NS and AES:

- **Spectral modeling (F-LSTM):** Near-end speech exhibits strong spectral correlations and harmonic structure due to its quasi-periodic nature in voiced segments [39]. By treating each time frame as a sequence of frequency bins, the bidirectional F-LSTM models the spectral envelope effectively, enabling separation between speech harmonics and the stochastic, broadband characteristics of environmental noise.

- **Temporal modeling (T-LSTM):** Acoustic echo is inherently governed by the room impulse response (RIR) and the convolutional relationship between far-end and microphone signals [1]. The T-LSTM operates along the time dimension, enabling the model to track echo evolution and capture residual components that persist after linear AEC.

Architectural Topology The bottleneck is designed to balance modeling capacity with the latency constraints of real-time DSP systems. It consists of a hierarchical stack of three F-T-LSTM layers, each extracting progressively more abstract latent representations.

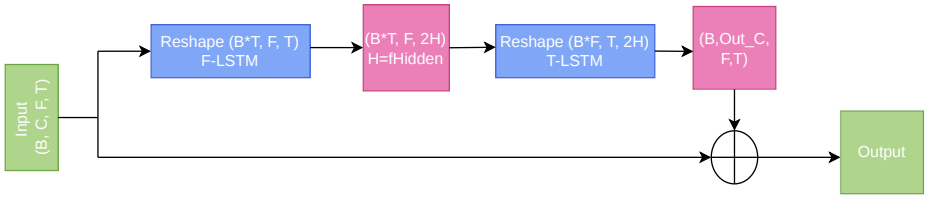


Figure 4.3: Structure of the proposed F-T-LSTM layer. The module first applies bidirectional recurrent modeling along the frequency axis, followed by temporal modeling along the time axis, enabling joint spectral–temporal representation learning for speech enhancement and AEC.

1. **Projection and fusion (Layer 1):** This layer acts as the interface between encoder and bottleneck. It projects the encoder output into a compressed latent space ($C = 128$), retaining only the most prominent features for interference suppression. The hidden dimensions are set to $f_{\text{hidden}} = 64$ and $t_{\text{hidden}} = 128$ for the F-LSTM and T-LSTM, respectively, reflecting their different modeling roles.

2. **Iterative refinement (Layers 2 and 3):** These layers refine the latent representation without further dimensional reduction. They focus on capturing long-range dependencies and late reverberation effects while maintaining a stable feature representation across time and frequency.

To ensure numerical stability and improve gradient flow, each F-T-LSTM block incorporates a residual connection. The output of the recurrent processing stage is added to the block input, enabling additive feature refinement while preserving information from earlier layers.

The asymmetric choice of hidden dimensions is motivated by the differing complexity of spectral and temporal modeling. The F-LSTM operates over local frequency structures within each frame, where harmonic relationships and spectral envelopes dominate the representation of speech signals [19].

In contrast, the T-LSTM models temporal evolution of the acoustic scene, including residual echo decay and late reverberation governed by the RIR, which persists over extended time spans [18]. These dynamics exhibit higher variability over time and therefore require a

more expressive latent representation.

Mathematical Formulation Let the input tensor be $\mathcal{X} \in R^{B\times C\times F\times T}$, where B is the batch size, C the number of channels, F the number of frequency bins, and T the number of temporal frames. The F-T-LSTM processes this representation through two sequential recurrent stages by explicitly reshaping the tensor along the frequency and temporal dimensions.

1. Frequency-Axis Modeling:

To model spectral dependencies, the input is first rearranged such that each time frame is treated as an independent sequence over frequency:

$$\mathcal{X}_f \in R^{(B\cdot T)\times F\times C}. \tag{4.25}$$

A bidirectional LSTM is then applied along the frequency axis:

$$H_f = \text{BiLSTM}_f(\mathcal{X}_f). \tag{4.26}$$

This allows the model to capture local and global spectral correlations within each time frame, leveraging the non-causal structure of the frequency domain representation.

2. Time-Axis Modeling:

The frequency-encoded representation is reshaped to enable temporal sequence modeling:

$$\mathcal{X}_t \in R^{(B\cdot F)\times T\times 2H}. \tag{4.27}$$

where $2H = 2f_{\text{hidden}}$ denotes the output dimensionality of the frequency LSTM it is 2 times the size due to the use of the bidirectional LSTM. A unidirectional LSTM is then applied along the temporal axis:

$$(H_t, h_t) = \text{LSTM}_t(\mathcal{X}_t, h_{t-1}). \tag{4.28}$$

where h_{t-1} denotes the recurrent hidden state propagated from the previous time step. The tuple (H_t, h_t) represents the output features and the updated hidden state, respectively, where h_t is passed to the next time step to maintain temporal memory.

This formulation ensures causal temporal modeling, allowing the network to maintain memory across time while processing sequential speech segments.

3. Reconstruction and Residual Learning:

The output is reshaped back to the original spectro-temporal layout:

$$H_{out} \in R^{B \times C_{out} \times F \times T}. \tag{4.29}$$

A residual connection with a 1×1 convolution ensures dimensional consistency:

$$Y = H_{out} + \mathcal{P}(\mathcal{X}). \tag{4.30}$$

where $\mathcal{P}$ is used to align channel dimensions when necessary. This residual formulation encourages the network to learn a corrective refinement of the input representation rather than a full transformation.

Causality and Stateful Temporal Modeling A key requirement of the proposed system is causality, as the target application is real-time AEC. Future frames are not available at inference time, so all enhancement must rely solely on current and past observations. This motivates the use of a unidirectional T-LSTM operating along the temporal axis, consistent with causal streaming constraints commonly adopted in real-time speech enhancement systems [69, 19].

In addition to causality, block-based processing introduces a limitation in temporal context, since each forward pass operates on short segments. To address this, the LSTM is implemented in a stateful manner, where hidden states are propagated across consecutive blocks instead of being reset. This form of state propagation is widely used in streaming

recurrent neural networks for speech processing tasks, where long-form dependencies must be maintained under chunk-wise inference constraints [70].

The mechanism allows the model to accumulate temporal information over long durations, effectively extending memory beyond a single processing window. Which is particularly important in AEC, where the echo signal is governed by the RIR, propagation delay, and multi-path reflections, which can introduce dependencies spanning hundreds of milliseconds [1].

A stateless formulation would restrict the model to short-term dependencies, reducing its ability to capture late reverberation and weakening overall cancellation performance. This stateful design is applied only to the temporal modeling stage. The F-LSTM remains stateless, as spectral modeling focuses on within-frame harmonic structure rather than long-range temporal evolution.

Overall, the combination of a unidirectional and stateful T-LSTM enables long-range temporal modeling while preserving strict causality, making the architecture suitable for real-time deployment in non-stationary acoustic environments.

Decoupled Frequency–Time Modeling The proposed bottleneck adopts a decoupled frequency–time modeling strategy, where spectral and temporal dependencies are treated as separate but complementary processes. The F-LSTM operates in a bidirectional manner across the frequency axis to exploit non-causal harmonic structure and spectral correlations within each frame. In contrast, the T-LSTM models temporal evolution in a strictly causal manner to satisfy real-time AEC constraints. This separation aligns with the distinct physical roles of spectral content and temporal dynamics in acoustic signal formation governed by the RIR, enabling more structured and efficient representation learning for residual echo and noise suppression [42].

Transition to Decoder The output of the F-T-LSTM bottleneck is a temporally enriched and spectrally structured latent representation in which both long-range dependencies and

local acoustic features have been modeled. This representation is passed to the decoder stage, where the enhanced time–frequency structure is progressively reconstructed to produce the final output.

Decoder Architecture

The decoder reconstructs the enhanced time–frequency representation from the bottleneck by progressively restoring resolution and integrating encoder features through skip connections. This combines global contextual information with fine spectral detail during reconstruction.

Unlike a strict inversion of the encoder, the decoder integrates bottleneck representations with multi-scale encoder features through skip connections. This ensures that global contextual information from the bottleneck is combined with high-resolution spectral details, preserving both structural consistency and fine acoustic detail during reconstruction.

The goal of this stage is therefore not only spatial upsampling, but also the generation of a perceptually coherent speech representation in which residual echo and noise artifacts are further suppressed during reconstruction.

TransGConv Blocks The decoder is built using TransGConv blocks, which serve as the upsampling counterpart to the encoder's gated convolutional modules. The internal structure of the TransGConv block is illustrated in Figure 4.4. Each block performs learned upsampling via transposed convolutions, followed by gated feature modulation. Similar to the encoder design, the input feature map is processed by parallel feature and gating branches, where the gating branch produces a sigmoid-activated mask that regulates information flow in a data-dependent manner.

This design is consistent with gated convolutional architectures used in deep feature refinement networks [62], where adaptive gating improves robustness in non-stationary acoustic conditions. In the decoder, this mechanism is particularly important for selectively reconstructing relevant spectral components while suppressing artifacts introduced during upsam-

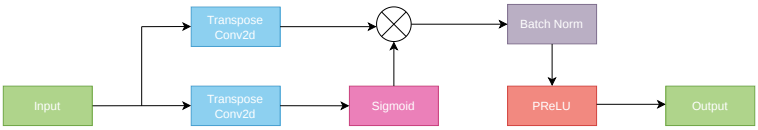


Figure 4.4: Structure of the TransGConv block used in the decoder. Transposed convolutions perform learned upsampling while a gating branch adaptively regulates feature flow during reconstruction.

pling.

While encoder layers uses compact 3×3 kernels for feature extraction, the decoder uses 4×4 kernels to reduce checkerboard artifacts and improve reconstruction smoothness [71]. This is especially relevant in encoder–decoder audio enhancement systems, where accurate interpolation of time–frequency representations is required.

Each TransGConv block uses a stride of 2, enabling symmetric $2\times$ upsampling that mirrors the encoder's downsampling path. Padding is set to 1 to maintain alignment with skip-connected feature maps and ensure consistent spatial dimensions across fusion stages.

Batch normalization is applied after each block to stabilize feature distributions and improve training convergence [65]. A PReLU activation is then used to introduce learnable non-linearity while preserving gradient flow in negative activation regimes [66].

Decoder Structure The decoder follows a predominantly symmetric expanding architecture that mirrors the encoder hierarchy. It is composed of a stack of TransGConv blocks that progressively increase time–frequency resolution while reducing channel dimensionality, as summarized in Table 4.2. The deepest encoder representation is first processed by the F-T-LSTM bottleneck and serves as the decoder input.

Unlike the original full U-Net-style design in [15], which employs a one-to-one correspondence between encoder and decoder stages with skip connections at every scale, the proposed architecture intentionally removes the final skip connection and its corresponding decoder stage. This modification is motivated by the observation that the F-T-LSTM bottleneck already captures sufficiently rich global context at the deepest representation, making the

lowest-resolution skip pathway partially redundant. Empirically, retaining the full set of skip connections yielded comparable performance while increasing architectural complexity, whereas the simplified design maintained performance while reducing redundancy and improving efficiency. A more detailed comparison is provided in the ablation study in Section 6.6.

Stage	**Input Ch.**	**Output Ch.**	**Stride**	**Padding**
L4 (Upsample)	$128 + 128$	128	2	1
L3 (Upsample)	$128 + 64$	64	2	1
L2 (Upsample)	$64 + 32$	32	2	1
Output Projection	32	2	1	0

Table 4.2: Decoder architecture consisting of transposed convolutional upsampling stages followed by a final 1×1 output projection layer producing the complex ratio mask representation.

Hierarchical Upsampling and Reconstruction Each decoding stage performs learned upsampling followed by feature refinement. For stages where a corresponding encoder feature map exists, skip connections are used to incorporate multi-scale information from the encoder.

This fusion mechanism allows reconstruction to be jointly informed by the long-range contextual representations learned by the F-T-LSTM bottleneck and the fine-grained spectral details preserved within the encoder feature maps.

Output Projection and Complex Ratio Mask Estimation The final decoder feature representation is obtained after the last upsampling and skip-fusion stage:

$$\mathcal{H}_{out} = \mathcal{H}_1. \tag{4.31}$$

where $\mathcal{H}_1$ denotes the output of the final TransGConv block following the third decoder stage, i.e., after progressive fusion with encoder skip connections at multiple scales.

The CRM is then estimated using a 1×1 convolutional projection:

$$\mathbf{Z} = \mathbf{W}_{out} * \mathcal{H}_{out}. \tag{4.32}$$

where $\mathbf{W}_{out}$ represents the learnable parameters of the final 1×1 convolution layer and $\mathbf{Z} \in R^{B\times 2\times F\times T}$ contains the predicted real and imaginary mask components. This formulation follows complex spectral modeling approaches commonly employed in phase-aware speech enhancement systems [57, 21].

A bounded hyperbolic tangent activation is then applied to stabilize the output range:

$$\hat{\mathbf{M}} = 2 \cdot \tanh(\mathbf{Z}). \tag{4.33}$$

where $\hat{\mathbf{M}} \in R^{B\times 2\times F\times T}$ denotes the estimated CRM. This operation constrains the predicted real and imaginary mask components to the interval $[-2, 2]$, reducing the likelihood of unstable spectral estimates and improving training stability, consistent with bounded-output formulations used in deep speech enhancement networks [57, 67].

The estimated CRM is applied to the noisy microphone spectrum $\mathbf{D}$ to obtain the enhanced complex spectrum:

$$\hat{\mathbf{Y}} = \hat{\mathbf{M}} \odot \mathbf{D}. \tag{4.34}$$

where $\mathbf{D}$ denotes the noisy microphone spectrum and $\odot$ represents element-wise complex multiplication in the real–imaginary representation of the spectrum.

Equivalently, this operation can be expressed in the time–frequency domain as:

$$\hat{Y}(f,t) = M(f,t) \cdot D(f,t). \tag{4.35}$$

where $M(f,t)$ and $D(f,t)$ denote the scalar time–frequency representations of $\hat{\mathbf{M}}$ and $\mathbf{D}$, respectively.

4.3.3 Output Formulation

Complex Ratio Mask Output and Signal Reconstruction

The model outputs a CRM, rather than directly estimating the enhanced time-domain signal as in [15]. The decoder outputs a two-channel representation corresponding to the real and imaginary components of a complex-valued mask applied to the microphone STFT [21].

Motivation for Complex Mask Estimation This formulation enforces a multiplicative correction of the observed spectrum rather than direct signal synthesis. Mask-based speech enhancement approaches have been shown to be more robust than direct spectral mapping methods, as they constrain the model to refine the input mixture rather than generate a full spectral representation from scratch [72, 57].

Compared to direct prediction approaches, this constraint reduces sensitivity to distributional mismatch between training and deployment conditions, improving generalization to unseen acoustic environments [57].

As a result, the underlying speech structure is better preserved while residual echo and noise components are selectively attenuated, leading to reduced perceptual distortion and improved reconstruction quality [21].

Block-Based Reconstruction and Overlap-Add Processing For real-time operation, processing is performed on overlapping time–frequency blocks. After applying the estimated CRM to each block, the enhanced signal is reconstructed via Inverse Short-Time Fourier Transform (ISTFT). The resulting time-domain blocks are combined using an overlap-add procedure with uniform weighting:

$$\hat{x}(n) = \frac{\sum_k x_k(n)}{\sum_k 1}. \tag{4.36}$$

where $x_k(n)$ denotes the reconstructed time-domain signal for the k-th block. The denominator corresponds to the number of overlapping contributions at each sample index.

This normalization ensures continuity across block boundaries and prevents amplitude scaling due to overlapping segments. Note that the analysis and synthesis windowing is handled internally by the STFT/ISTFT pair, while the block-level reconstruction applies uniform averaging rather than an explicit synthesis window [19].

Output Stabilization A global amplitude normalization step is applied after reconstruction to maintain a consistent dynamic range relative to the input signal. This prevents clipping and ensures the enhanced output remains physically consistent under varying acoustic conditions.

Together, these constraints ensure stable reconstruction and robust performance in non-stationary AEC scenarios.

4.4 Loss Functions

The proposed post-filter is trained using a multi-component loss function designed to jointly optimize spectral accuracy, phase consistency, and time-domain perceptual speech quality. In the context of AEC, the training objective must carefully balance competing requirements: achieving high attenuation of residual echo and background noise during single-talk and double-talk intervals, while preserving the acoustic integrity of the near-end speech signal. Prior work in speech enhancement has demonstrated that a single objective function is often insufficient to simultaneously optimize these distinct acoustic parameters, motivating the use of composite perceptual and spectral loss formulations [73, 74].

Given that the model estimates a CRM, the loss function is structured to guide the network toward learning an effective time–frequency masking strategy rather than directly synthesizing the waveform from unconstrained features.

The aggregate loss formulation is conceptually inspired by [15], but is adapted to reflect the proposed network architecture and its role as a post-filter within an automotive AEC pipeline. Several alternative loss configurations were explored during development, the

impacts of which are detailed in the ablation study presented in Section 6.6.

The loss computation begins by reconstructing the estimated near-end speech spectrum in the time–frequency domain using the predicted complex mask.

4.4.1 Signal Reconstruction for Loss Computation

Following the output formulation described in Section 4.3, the predicted CRM is applied directly to the microphone STFT representation to obtain a complex estimate of the clean near-end speech. The reconstructed spectrum is defined as:

$$\hat{S}[f,t] = M[f,t] \cdot D[f,t]. \tag{4.37}$$

where $\hat{S}[f,t]$ denotes the estimated clean speech in the time-frequency domain, $M[f,t]$ is the predicted complex ratio mask, and $D[f,t]$ is the noisy microphone STFT signal. The corresponding ground-truth target signal is denoted as $S[f,t]$.

From these complex quantities, the respective magnitude spectral representations are computed via the absolute operator:

$$|\hat{S}[f,t]| = \sqrt{\Re(\hat{S}[f,t])^2 + \Im(\hat{S}[f,t])^2}, \quad |S[f,t]| = \sqrt{\Re(S[f,t])^2 + \Im(S[f,t])^2}. \tag{4.38}$$

which serve as the mathematical basis for subsequent magnitude-based loss terms [19].

4.4.2 Voice Activity-Aware Weighting

To improve training stability and ensure high echo attenuation during periods of near-end silence, a voice activity-aware weighting scheme is integrated into the spectral objective functions. Unlike configurations that down-weight silent regions to avoid background noise optimization, the proposed framework intensifies gradient penalties during near-end silence to aggressively eliminate residual echo leakage when no masking speech is present.

A thresholded frame-level weight is constructed based on the average energy of the target near-end speech signal. The average spectral energy for a given frame index t is computed from the target magnitude spectrum as:

$$E[t] = \frac{1}{F} \sum_{f=0}^{F-1} |S[f,t]|^2. \tag{4.39}$$

where F represents the total number of frequency bins. The frame-level energy is mapped to a regional weighting mask $W[t]$ via an empirical step function:

$$W[t] = \begin{cases} \gamma_{\text{silence}}, & E[t] < \tau \\ 1.0, & \text{otherwise} \end{cases}. \tag{4.40}$$

where τ is a predefined energy threshold set to 10^{-5}, and $\gamma_{\text{silence}} > 1.0$ acts as an optimization multiplier (empirically set to 2.0) applied to up-weight the loss during periods of near-end silence. This weighting factor ensures that the network is heavily penalized for any echo bleeding through when the near-end speaker is inactive.

This frame-level weight matrix $W[t]$ is subsequently applied as a multiplicative scaling factor to the individual magnitude and complex loss components, ensuring that the total optimization trajectory is strictly directed toward speech preservation during double-talk and maximum suppression during single-talk echo intervals [75, 76].

4.4.3 Magnitude Loss with Asymmetric Penalty

To enforce precise spectral envelope reconstruction while maintaining perceptual audio alignment, a power law compressed magnitude loss is implemented. Power-law compression compresses the high dynamic range of the acoustic spectrogram, mimicking human loudness perception and preventing high-amplitude components from dominating the gradients over low-amplitude speech cues [77].

The power-law compressed spectral magnitudes are defined as:

$$\tilde{\hat{S}}[f,t] = |\hat{S}[f,t]|^p, \quad \tilde{S}[f,t] = |S[f,t]|^p. \tag{4.41}$$

where the compression factor is empirically set to $p = 0.3$. The resulting compressed magnitude estimation error is given by:

$$e_{\text{mag}}[f,t] = \tilde{\hat{S}}[f,t] - \tilde{S}[f,t]. \tag{4.42}$$

To reflect the asymmetric perceptual sensitivity inherent to telecommunication environments, a directional penalty framework is introduced [78]. Over-estimation errors ($e_{\text{mag}}[f,t] > 0$), which correspond to audible residual echo leakage, are penalized more severely than under-estimation errors ($e_{\text{mag}}[f,t] < 0$), which typically manifest as less objectionable near-end speech attenuation. The asymmetric frame-wise loss is defined as:

$$\mathcal{L}_{\text{mag}}[f,t] = \begin{cases} \frac{1}{2} e_{\text{mag}}[f,t]^2, & e_{\text{mag}}[f,t] < 0 \\ \alpha_{\text{asym}}\, e_{\text{mag}}[f,t]^2, & e_{\text{mag}}[f,t] \geq 0 \end{cases}. \tag{4.43}$$

where α_{asym} controls the severity of the echo leakage penalty and is set to 4.0.

The aggregate magnitude loss is computed as the total mean expectation across the batch, frequency bins, and frames, scaled by the voice activity mask:

$$\mathcal{L}_{\text{mag, total}} = \frac{1}{B \cdot F \cdot T} \sum_{b=0}^{B-1} \sum_{t=0}^{T-1} \sum_{f=0}^{F-1} W[b,t] \cdot \mathcal{L}_{\text{mag}}[b,f,t]. \tag{4.44}$$

where B represents the training batch size.

4.4.4 Power-Law Compressed Phase-Aware Loss

While the magnitude loss optimizes spectral envelope definitions, a phase-aware complex-domain loss is concurrently executed to preserve phase phase consistency and harmonic alignments. Rather than isolating phase angles, which exhibit severe mathematical discontinuities at $\pm\pi$, the formulation extends the power-law compression scheme directly into the complex plane [15].

The estimated and target STFT representations are represented as phase-vector components scaled by their compressed magnitudes using the identical factor $p = 0.3$ established in Equation 4.41:

$$\hat{\mathbf{Z}}[f,t] = |\hat{S}[f,t]|^p \cdot \frac{\hat{S}[f,t]}{|\hat{S}[f,t]|}, \quad \mathbf{Z}[f,t] = |S[f,t]|^p \cdot \frac{S[f,t]}{|S[f,t]|}. \tag{4.45}$$

The phase-aware distance metric is subsequently computed across the orthogonal real ($\Re$) and imaginary ($\Im$) components using a mean squared tracking format:

$$\mathcal{L}_{\text{cplx}}[f,t] = \Big(\Re(\hat{\mathbf{Z}}[f,t]) - \Re(\mathbf{Z}[f,t])\Big)^2 + \Big(\Im(\hat{\mathbf{Z}}[f,t]) - \Im(\mathbf{Z}[f,t])\Big)^2. \tag{4.46}$$

By operating inside this compressed complex space, the optimization jointly constrains phase variance alongside structural spectral transforms. Mirroring the magnitude pipeline, the aggregate complex loss is computed as:

$$\mathcal{L}_{\text{cplx, total}} = \frac{1}{B \cdot F \cdot T} \sum_{b=0}^{B-1} \sum_{t=0}^{T-1} \sum_{f=0}^{F-1} W[b,t] \cdot \mathcal{L}_{\text{cplx}}[b,f,t]. \tag{4.47}$$

4.4.5 SI-SDR Loss

To optimize global waveform structure directly in the time domain, a scale-invariant signal-to-distortion ratio loss is incorporated. Unlike spectral-domain metrics, SI-SDR evaluates the similarity between the reconstructed estimated waveform and the reference clean waveform

in the time domain, making it effective at penalizing overall reconstruction distortions [79].

Let $\hat{s}[n]$ and $s[n]$ denote the estimated and target time-domain waveforms, respectively, obtained via inverse STFT of the reconstructed spectra.

To ensure scale invariance and remove Direct Current (DC) offset bias, both signals are zero-mean normalized:

$$\tilde{\hat{s}}[n] = \hat{s}[n] - \mu_{\hat{s}}, \quad \tilde{s}[n] = s[n] - \mu_s. \tag{4.48}$$

where μ denotes the mean value over the temporal dimension.

The target projection is defined as:

$$s_{\text{target}} = \frac{\langle \tilde{\hat{s}}, \tilde{s} \rangle}{\|\tilde{s}\|^2} \tilde{s}. \tag{4.49}$$

and the residual error is:

$$e_{\text{noise}} = \tilde{\hat{s}} - s_{\text{target}}. \tag{4.50}$$

The SI-SDR metric is then computed as:

$$\text{SI-SDR} = 10 \log_{10} \left(\frac{\|s_{\text{target}}\|^2}{\|e_{\text{noise}}\|^2} \right). \tag{4.51}$$

To stabilize training, the loss is defined as the negative clipped SI-SDR:

$$\mathcal{L}_{\text{SI-SDR}} = -\text{clamp}(\text{SI-SDR}, -30, 50). \tag{4.52}$$

4.4.6 Complex Mask Regularization (Bias Term)

To stabilize training parameters and discourage unconstrained mask boundaries, a weak regularization constraint is applied directly to the raw predicted CRM coefficients. This regularization term implicitly biases the network toward an identity mapping ($M[f,t] = 1 + 0j$) when acoustic conditions do not present clear evidence of interference, thereby avoiding aggressive over-suppression artifacts.

Decomposing the predicted complex mask into its real and imaginary tracking channels, $M[f,t] = M_{\mathrm{R}}[f,t] + jM_{\mathrm{I}}[f,t]$, the mask bias loss is computed uniformly over the grid as:

$$\mathcal{L}_{\text{bias}} = \frac{1}{B \cdot F \cdot T} \sum_{b=0}^{B-1} \sum_{t=0}^{T-1} \sum_{f=0}^{F-1} \left[\left(M_{\mathrm{R}}[b,f,t] - 1.0 \right)^2 + \left(M_{\mathrm{I}}[b,f,t] \right)^2 \right]. \tag{4.53}$$

This penalty directly discourages unnecessary phase shifting or magnitude scaling during intervals characterized by highly clean near-end speech or general silence, preventing the degradation of acoustic transparency [21].

4.4.7 Overall Loss Function

The final training objective is formulated as a linear combination of the spectral, complex-domain, time-domain, and structural regularization terms. This multi-component structure allows the model to jointly prioritize reconstruction fidelity, phase alignment, and tracking stability within the AEC framework.

The total objective function minimized by the optimizer is defined as:

$$\mathcal{L}_{\text{total}} = \alpha \, \mathcal{L}_{\text{cplx, total}} + (1 - \alpha) \, \mathcal{L}_{\text{mag, total}} + \lambda_{\text{SI-SDR}} \, \mathcal{L}_{\text{SI-SDR}} + \lambda_{\text{bias}} \, \mathcal{L}_{\text{bias}}. \tag{4.54}$$

where the hyperparameters match the implementation values:

- $\alpha = 0.5$, balancing the optimization focus evenly between magnitude envelopes and complex phase vectors,
- $\lambda_{\text{SI-SDR}} = 0.2$, assigning a moderate scaling weight to time-domain waveform alignment,
- $\lambda_{\text{bias}} = 0.01$, acting as a highly attenuated stabilizing prior to encourage identity mappings when interference is absent.

Together, these components establish a highly balanced optimization routine engineered

specifically for non-stationary real-world automotive environments, minimizing speech tracking distortions while enforcing extreme attenuation across complex echo environments.

4.5 Baseline System

Due to disclosure constraints from our industry partner, only a high-level description of the baseline system is provided. The baseline follows a conventional modular speech enhancement pipeline commonly used in classical AEC and speech processing systems. It is designed according to a cascaded architecture in which different sources of degradation are addressed in separate, sequential processing stages.

4.5.1 Traditional Processing Pipeline

The baseline system follows a cascaded processing structure in which AEC is applied first, followed by AES and NS. This sequential design reflects a modular signal processing paradigm in which each stage is responsible for mitigating a specific class of distortion in the microphone signal. The AEC stage is the same one used as the input to the proposed hybrid model in this paper.

In this pipeline, AES operates on the partially processed signal produced by the AEC stage and aims to further attenuate residual echo components that are not fully removed during linear cancellation. This stage typically applies time–frequency domain attenuation based on statistical or heuristic estimates of residual interference. In this case it is done using a Wiener Filter.

The final stage, NS, targets non-echo background noise such as stationary or slowly varying environmental noise sources. This stage refines the output of the cascade by applying spectral gain-based suppression, improving overall speech intelligibility and perceptual quality. In practice, NS modules are often designed using classical statistical estimators that adaptively compute attenuation factors in the time–frequency domain.

Together, AES and NS provide complementary suppression of residual interference and background noise, completing the traditional three-stage speech enhancement pipeline.

This cascaded structure reflects the traditional design philosophy in speech enhancement systems, where each component is optimized independently for a specific type of distortion.

4.5.2 Wiener-Based Suppression

The AES and NS stages in the baseline system are based on classical Wiener filtering principles. These methods operate in the time–frequency domain by estimating a soft spectral gain function that attenuates interference components while preserving the underlying speech structure.

In a joint acoustic echo and noise control configuration, the error signal $E[f,t]$ captured at the output of the linear AEC contains target near-end speech $S[f,t]$, ambient background noise $N[f,t]$, and a residual echo component $B[f,t]$ arising from filter misalignment, room impulse response truncation, or system nonlinearities [80]:

$$E[f,t] = S[f,t] + N[f,t] + B[f,t]. \tag{4.55}$$

where f denotes the discrete frequency bin and t represents the short-time frame index. Assuming mutual statistical independence between the speech, ambient noise, and far-end echo components, the total interference PSD $\Phi_{nn}[f,t]$ can be modeled as the linear summation of the individual residual echo and background noise power spectra [80]:

$$\Phi_{nn}[f,t] = \Phi_{BB}[f,t] + \Phi_{NN}[f,t]. \tag{4.56}$$

where $\Phi_{BB}[f,t]$ represents the residual echo PSD, and $\Phi_{NN}[f,t]$ denotes the background noise PSD.

The Wiener filter is derived as the MMSE estimator under the assumption of uncorrelated speech and interference components. This results in the well-known spectral gain

formulation:

$$G[f,t] = \frac{\Phi_{ss}[f,t]}{\Phi_{ss}[f,t] + \Phi_{nn}[f,t]}. \tag{4.57}$$

where $\Phi_{ss}[f,t]$ denote the PSD estimates of the target speech.

The resulting gain is applied multiplicatively to the input spectrum, effectively acting as a soft mask that suppresses low signal-to-interference regions while preserving higher confidence speech components. This formulation is widely used in practical speech enhancement systems due to its computational efficiency and stability.

However, its performance depends strongly on the accuracy of PSD estimation and the validity of the underlying statistical assumptions. To achieve robust attenuation in highly dynamic environments, specialized classical estimators must be implemented to track the components $\Phi_{BB}[f,t]$ and $\Phi_{NN}[f,t]$ independently.

Residual Echo Power Spectral Density Estimation

The estimation of the residual echo power spectrum $\Phi_{BB}[f,t]$ follows the coherence-based framework established by Enzner et al. [80]. This approach circumvents explicit acoustic path modeling by estimating the relationship between the far-end excitation signal $X[f,t]$ and the linear AEC's error spectrum $E[f,t]$ in the statistical domain. Under full-duplex conditions, the system evaluates the linear coupling between these two signals via the magnitude-squared coherence function $C_{XE}[f,t]$, defined as:

$$C_{XE}[f,t] = \frac{|\Phi_{XE}[f,t]|^2}{\Phi_{XX}[f,t]\Phi_{EE}[f,t]}. \tag{4.58}$$

where $\Phi_{XE}[f,t]$ represents the cross-PSD between the far-end and error signals, while $\Phi_{XX}[f,t]$ and $\Phi_{EE}[f,t]$ denote their respective auto-PSDs [80].

The coherence function serves as a frequency-dependent indicator of the proportion of the error signal power that is linearly correlated with the far-end speaker. Then, the residual echo power spectrum is recovered by weighting the total auto-PSD of the error signal with

this coherence measure [80]:

$$\Phi_{BB}[f,t] = C_{XE}[f,t]\Phi_{EE}[f,t]. \tag{4.59}$$

In practical systems, the required auto- and cross-power spectral densities are estimated using short-term statistical averaging over time. However, this general estimation strategy is subject to severe physical and structural limitations in real-world acoustic environments. Because physical enclosures exhibit long reverberation times that exceed the brief time windows mandated by low-latency constraint guidelines, short-term spectral analysis fails to capture the full temporal dispersion of the echo path. This leads to an underestimation of late reverberant echo components [80]. Additionally, during periods of simultaneous near-end and far-end speech (double-talk), independent local signals introduce erroneous short-term cross-correlations that artificially inflate the coherence estimate, resulting in overestimation bias and subsequent clipping of the near-end speech [80]. Bypassing these fundamental tracking errors motivates the transition toward data-driven models that can inherently map these non-linear conditions without relying on stationary coherence assumptions.

Background Noise Power Spectral Density Estimation

The baseline system must also track the non-stationary background noise spectrum $\Phi_{NN}[f,t]$. This is done using an augmented version of the minimum statistics approach formulated by Martin [26]. This methodology operates without an explicit VAD, thereby avoiding the threshold-tuning challenges and frequent near-end speech clipping associated with hard-decision VAD schemes under adverse SNR conditions. The minimum statistics technique leverages the physical observation that even during active speech periods, the short-term power of a noisy mixture continuously decays to the background noise floor during brief pauses, word boundaries, or syllabic intervals.

Instead of employing binary voice activity classifications, the algorithm tracks the noise

floor by continually searching for spectral minima within a finite sliding window of length D frames. To prevent the highly dynamic, high-energy excursions of active speech from bleeding into and corrupting the noise estimate, the PSD of the error signal is smoothed using a time-varying, frequency-dependent parameter $\alpha[f,t]$:

$$P[f,t] = \alpha[f,t]P[f,t-1] + (1-\alpha[f,t])|E[f,t]|^2. \tag{4.60}$$

where $P[f,t]$ is the smoothed power spectrum, and $|E[f,t]|^2$ represents the instantaneous power spectrum of the linear AEC error signal. This smoothing parameter is optimized based on a conditional mean-square error criterion that monitors a posteriori power variations. This optimization ensures that the smoothing factor dynamically approaches zero during speech transients and returns to its maximum boundary during speech absences to minimize estimation variance. Based on these tracked minima, the final background noise power spectrum is formulated as:

$$\Phi_{NN}[f,t] = B_{\min}[f,t] \cdot \min_{\tau \in \{0,\ldots,D-1\}} P[f,t-\tau]. \tag{4.61}$$

where $B_{\min}[f,t]$ represents the adaptive bias correction factor.

While computationally efficient and highly stable, this classical tracking framework exhibits inherent structural trade-offs. The length of the sliding window D represents a critical compromise: it must be long enough to bridge extended speech utterances without capturing the speech envelope, yet short enough to track rapid changes in non-stationary noise fields. In highly dynamic or non-stationary acoustic environments, tracking delays within this sliding window, combined with imperfections in the adaptive bias correction under tracking lag, manifest as local noise estimation errors or musical noise artifacts. These tracking limitations lead towards the motivation for transitioning toward data-driven deep learning architectures, which bypass explicit parametric window constraints by jointly learning the continuous, multi-scale mapping of complex speech and non-stationary noise topologies.

4.5.3 Limitations of the Baseline

Although the baseline system is robust and widely deployed in practical speech enhancement applications, its modular design introduces several architectural limitations when handling complex, real-world acoustic scenes.

First, the cascaded structure results in sequential error propagation across the processing chain. Because the linear AEC, AES, and NS modules are executed in series, any modeling errors or transient artifacts produced by an upstream stage are passed directly to subsequent blocks. For example, uncancelled residual echo spikes exiting the linear AEC can corrupt the downstream noise estimator, causing it to misinterpret echo as non-stationary background noise. Since each module operates independently without a joint feedback mechanism, the system lacks a method to correct these compound distortions downstream.

Second, each processing stage is optimized in isolation using independent, statistically derived objectives. This isolated optimization assumes strict statistical boundaries, such as mutual independence and stationarity between speech, echo, and noise components. In highly dynamic or interactive full-duplex scenarios, these classical assumptions frequently collapse. The decoupled design prevents the system from adapting to the complex, non-linear interactions between loudspeaker distortion, room reverberation, and ambient noise fields, leading to localized performance degradation.

Third, the multiplicative soft-masking framework of the classical Wiener filter imposes a fundamental theoretical bound on signal separation. Even if perfect PSD tracking could be achieved, applying a real-valued, phase-invariant gain to a joint short-time spectrum inherently introduces an adversarial trade-off between suppression depth and target speech preservation. In challenging scenarios like double-talk, this linear gain function inevitably causes a choice between localized interference leakage or clipping and perceptual distortion of the near-end speech.

Finally, the lack of end-to-end optimization prevents the system from automatically learning multi-scale, task-specific representations that minimize global distortion. Rather than

jointly balancing echo and noise reduction against speech naturalness, the baseline relies entirely on the hand-tuned alignment of discrete algorithms. This fundamental constraint motivates the transition toward a unified, data-driven framework, where residual artifacts from the linear AEC are directly refined by a deep learning-based post-filter optimized via a comprehensive multi-domain loss function.

To evaluate this proposed learning-based framework against the classical baseline under controlled conditions, the following chapter details the experimental setup, including training configurations, dataset synthesis, and objective performance metrics.

Chapter 5

Experimental Setup

This chapter presents the experimental setup used to train and evaluate the proposed hybrid AEC and NS system. The framework combines a linear AEC front-end with a deep learning–based post-filter, and its performance is assessed under controlled and reproducible conditions.

The training pipeline is first described, including the dataset composition, preprocessing steps, and model training procedure used to learn the CRM. This is followed by a description of the test dataset and evaluation scenarios used to assess system performance.

Evaluation is conducted across three acoustic conditions: near-end single-talk, far-end single-talk, and double-talk. These scenarios are designed to probe key system capabilities, including speech preservation, echo suppression, and robustness under overlapping speech conditions. Finally, objective evaluation metrics are introduced to provide a consistent and comprehensive basis for comparison across all scenarios.

5.1 Training Dataset

The proposed hybrid AEC and NS system is trained using the synthetic dataset from the ICASSP AEC Challenge [12]. This dataset is designed to mimic realistic hands-free communication scenarios while still providing access to clean reference signals required for supervised

learning.

The dataset consists of 10,000 synthetic scenarios, each capturing a different acoustic condition. Each scenario includes a combination of far-end speech, near-end speech, echo, background noise, and nonlinear distortions. For every scenario, four signals are provided: the far-end speech, the generated echo, the near-end speech, and the resulting microphone signal. This structure enables supervised training, since the desired target signal (near-end speech) is provided.

In this work, only the synthetic training data of the dataset is used. This choice is motivated by the availability of clean near-end speech, which is required to compute the CRM training target. In contrast, real recordings do not provide access to isolated ground truth signals, making supervised training more difficult.

As illustrated in Figure 5.1, each synthetic training example is generated by combining far-end speech, near-end speech, room acoustics, background noise, and nonlinear distortions to emulate realistic hands-free communication conditions while preserving access to clean reference signals for supervised learning.

The underlying speech data is sourced from the LibriVox corpus, utilizing the clean and noisy subsets originally compiled for the Deep Noise Suppression challenge framework [81]. These high-quality audiobook segments were verified using the ITU-T P.808 crowdsourcing evaluation methodology [82], ensuring an initial baseline MOS typically ranging from 4.3 to 5.0. To better reflect real-world acoustic diversity, the non-stationary background noise environments were generated by mixing these clean speech clips with noise samples drawn from AudioSet [83], Freesound [84], and the Diverse Environments Multi-channel Acoustic Noise Database (DEMAND) database [85]. The SNR is uniformly sampled between 0 dB and 40 dB, allowing the model to see a wide range of noise levels during training.

Each training example is created by randomly selecting different speakers for the far-end and near-end signals. A 10-second segment is used for the far-end, while a longer near-end segment is selected and then truncated or zero-padded to match this duration. The speech

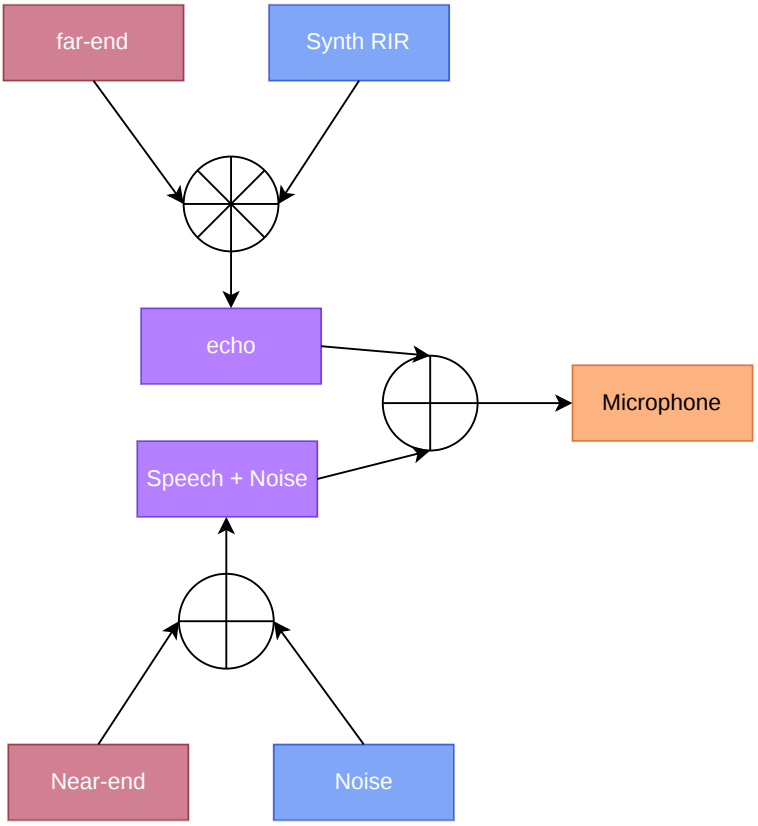


Figure 5.1: Overview of the synthetic data generation process used in the ICASSP AEC Challenge dataset. Far-end speech is convolved with a synthetic RIR to generate acoustic echo, while near-end speech and optional background noise are combined to create realistic microphone mixtures. The clean near-end speech is retained as the ground-truth target for supervised training.

activity within the selected segments is varied such that the same 10-second scenario can contain periods of near-end single-talk, far-end single-talk, and double-talk. Thus, these acoustic conditions are not represented by separate files; rather, the speech activity changes over time within each scenario. This provides the model with transitions between different operating conditions during training while maintaining a consistent signal duration and alignment.

Echo is simulated by convolving the far-end signal with randomly selected RIRs, with Reverberation Time (RT60) values ranging from 200 ms to 1200 ms. This allows the model

to experience both mildly and highly reverberant environments. In addition, nonlinear distortions are applied in 80% of the cases to mimic loudspeaker effects. These include clipping, sigmoidal transformations, and learned nonlinear functions, which make the task more realistic and challenging.

The final microphone signal is formed by mixing the echo with the near-end speech at a signal-to-echo ratio between -10 dB and 10 dB. In 50% of the cases, both signals are taken from the noisy dataset, meaning that echo and background noise are present at the same time. This encourages the model to handle echo cancellation and noise suppression jointly.

Overall, this dataset provides a controlled yet diverse training environment that captures the key challenges of real-world AEC systems, including double-talk, noise, reverberation, and nonlinearities. At the same time, access to clean reference signals makes it well suited for supervised learning, while the variability in conditions promotes strong generalization. The following section describes the preprocessing steps used to convert these signals into the time–frequency representations required for model training.

5.2 Training Preprocessing

The preprocessing pipeline used during training follows the signal representation introduced in Section 4.3. In this stage, the input signals are converted into time–frequency representations and organized into a structured input tensor suitable for model training.

For each training example, three signals are used as inputs to the model: the microphone signal, the echo estimate, and the output of the linear AEC system. These signals are converted to their complex STFT representations and separated into real and imaginary components. Each component is independently normalized to zero mean and unit variance to improve numerical stability and training convergence.

The normalized features are stacked along the channel dimension to form a six-channel

input tensor:

$$\mathbf{X} = [D_r, D_i, Y_r, Y_i, E_r, E_i]. \tag{5.1}$$

where D denotes the microphone signal, Y the echo estimate, and E the linear AEC output in the time-frequency domain. This representation provides the model with both the raw mixture and intermediate echo suppression information.

The training target is defined as a CRM computed between the microphone signal and the clean near-end speech. Given the complex STFT of the microphone signal $D = D_r + jD_i$ and the near-end speech $S = S_r + jS_i$, the real and imaginary components of the mask are given by:

$$M_r = \frac{S_r D_r + S_i D_i}{D_r^2 + D_i^2 + \epsilon}, \tag{5.2}$$

$$M_i = \frac{S_i D_r - S_r D_i}{D_r^2 + D_i^2 + \epsilon}. \tag{5.3}$$

where ϵ is a small constant for numerical stability. This formulation allows the model to jointly learn magnitude and phase corrections.

To enable efficient training on long sequences, the input features are processed in fixed-length time chunks. Consecutive segments are processed sequentially while maintaining the temporal hidden states between chunks, allowing the model to capture long-term temporal dependencies across the full utterance. Gradients are truncated at chunk boundaries to reduce memory consumption and ensure stable optimization during training.

During training, the model predicts a CRM for each input chunk, which is applied to the corresponding microphone STFT. Signal reconstruction is handled within the loss function, allowing optimization to be performed implicitly in the time domain.

Overall, this preprocessing pipeline prepares the input data for the proposed model by extracting features from both the linear AEC output and the microphone signal, then transforming them into normalized time-frequency representations. The resulting feature tensor provides a compact and informative description of the acoustic scene and serves as the input

to the training procedure described in the following section.

5.3 Training Procedure

The model is trained using the preprocessing pipeline described in Section 5.2 and the multi-component loss function defined in Section 4.4. This section outlines the practical training configuration and implementation details.

5.3.1 Optimization Setup

The model is optimized using the AdamW optimizer with a learning rate of 2.5×10^{-4} and a weight decay of 10^{-4}. AdamW is chosen due to its decoupled weight decay formulation, which has been shown to improve generalization and training stability compared to standard Adam, particularly in deep neural networks with large parameter counts [86]. In the context of speech enhancement and AEC, adaptive optimizers such as AdamW are commonly used due to their robustness to varying gradient scales in time–frequency representations [10].

To improve convergence stability, a learning rate schedule combining a manually implemented linear warmup and cosine annealing is used. During the first 10 epochs, the learning rate is linearly increased from 0.1 $\times$ base learning rate to the base learning rate, reducing instability during early optimization when model parameters are randomly initialized [87]. After the warmup phase, a cosine annealing schedule (implemented using PyTorch's `CosineAnnealingLR`) is applied over the remaining training epochs, gradually decaying the learning rate toward a minimum value of 10^{-6}. This allows for large parameter updates during early training and finer updates in later stages, improving overall convergence behavior [87].

Gradient norm clipping with a maximum norm of 5.0 is applied during training to improve numerical stability. This is particularly important given the recurrent structure of the model and the use of truncated backpropagation through time over long sequences, where exploding

gradients may occur. Gradient clipping helps constrain parameter updates and is a standard technique for stabilizing training in sequence modeling and recurrent neural network-based architectures [88].

5.3.2 Loss Configuration

Training is performed using the composite loss function described in Section 4.4. The final configuration uses a power-law compression factor of $p = 0.3$, an asymmetric penalty coefficient of $\alpha_{\text{asym}} = 5.0$, and a weighting factor of $\alpha = 0.5$ to balance magnitude and complex-domain losses.

The SI-SDR term is included with a weighting of 0.2 and is computed in the time domain using reconstructed waveforms. To ensure stable optimization, the SI-SDR term is applied only when sufficient temporal context is available. A small bias regularization term is also included to encourage identity mask behavior in low-interference regions.

In addition, a voice activity-aware weighting mechanism is applied during training, with a weighting factor of 2.0 used to emphasize echo suppression during low-energy regions.

5.3.3 Batching and Data Splitting

Training is performed using a batch size of 8. A total of 10,000 samples from the synthetic dataset are used, with a held-out subset reserved for validation. The validation set is used for parameter updates with a validation loss. While alternative stopping criteria based on perceptual metrics such as SI-SDR and ERLE were evaluated, they were found to introduce instability during training and did not consistently correlate with improved generalization performance. Therefore, model selection is based on the aggregated validation loss, which provides a more stable optimization signal under the proposed multi-component loss function.

Different batch sizes and learning rates were explored during preliminary experiments. Smaller batch sizes resulted in higher gradient variance and less stable convergence, while larger batch sizes reduced stochasticity and negatively impacted generalization under highly

non-stationary conditions such as double-talk. A batch size of 8 was found to provide the best trade-off between optimization stability and generalization performance.

This is consistent with common practice in STFT-based speech processing models, where moderate batch sizes are typically preferred due to the high dimensionality and strong temporal correlation of spectral features. Similar configurations have been reported in related speech enhancement and AEC systems, which also observe improved generalization using relatively small to moderate batch sizes [89].

The selected learning rate was determined empirically to ensure stable convergence without excessively slowing training. In practice, the choice of learning rate is closely coupled with batch size, as smaller batches introduce higher gradient variance and therefore typically require more conservative learning rates to maintain stability, whereas larger batches permit more aggressive updates due to smoother gradient estimates [90]. The chosen configuration reflects a balance between these factors, aligning the learning rate with the moderate batch size used in this work.

5.3.4 Chunk-wise Training Strategy

The model is trained on full continuous utterances that are processed sequentially in short temporal segments. During training, each utterance is divided into fixed-length segments of 16 frames, which are fed to the network in temporal order.

The recurrent components maintain their internal states across segments, allowing information to propagate through the full utterance and enabling the model to learn long-range temporal dependencies relevant to acoustic echo dynamics.

For computational efficiency, gradients are computed locally within each segment, while the hidden states are passed forward without interruption. This enables stable optimization while preserving the temporal continuity required for streaming operation.

5.3.5 Training Loop and Convergence

At each iteration, the model predicts a CRM for the current input chunk. The predicted mask is applied to the corresponding microphone STFT, and the loss is computed in the spectral domain using a multi-component objective that combines complex reconstruction, magnitude consistency, and time-domain constraints via the SI-SDR term.

Early stopping is applied with a patience of 8 epochs based on the validation loss. The model achieving the lowest validation loss is saved during training and used for final evaluation. Checkpointing is implemented to allow training to resume from intermediate states. Both model parameters and optimizer states are stored to ensure continuity across training sessions.

While alternative stopping criteria based on perceptual metrics such as SI-SDR and ERLE were investigated, they were found to produce unstable convergence behavior and inconsistent model selection. As a result, validation loss was adopted as the primary criterion for both early stopping and checkpoint selection, providing a more stable and reliable training signal under the proposed multi-objective loss formulation.

5.3.6 Validation Strategy

Model performance during training is evaluated on a held-out validation set using both the proposed validation loss and the ERLE metric [1]. The validation loss is used as the primary criterion for model selection and early stopping, while ERLE serves as an additional time-domain performance indicator.

During validation, the model processes full utterances using the same chunk-wise inference strategy as in training. The predicted masks are applied in the STFT domain, and the enhanced signal is reconstructed using Overlap-Add (OLA) followed by ISTFT.

Although ERLE is not directly optimized during training or used for checkpoint selection, it provides an additional measure of residual echo suppression in the time domain and is used to qualitatively assess model behaviour across epochs.

5.3.7 Implementation Details

All models are implemented in PyTorch and trained on Graphics Processing Unit (GPU)-enabled hardware when available. Device selection is handled dynamically to support different execution environments, allowing the same codebase to run on both Central Processing Unit (CPU) and GPU environments.

Training was performed on an AMD Radeon RX 7900 XT GPU. Due to hardware compatibility requirements, training was conducted on a Linux-based system using the ROCm software stack to enable GPU acceleration with PyTorch. This configuration allows efficient execution of large-scale time-frequency domain operations, particularly the STFT-based processing and chunk-wise recurrent computations used in the proposed model.

The final configuration presented in this work was selected based on empirical performance. Variations in optimizer settings, loss weighting, and input configurations are further explored in the ablation study presented in Section 6.6.

To evaluate the effectiveness and generalization capability of the proposed system, the following section introduces the test dataset and evaluation setup.

5.4 Test Dataset

The evaluation of the proposed system is conducted using the official test dataset provided as part of the ICASSP AEC Challenge [12]. This dataset is designed to reflect realistic acoustic conditions encountered in hands-free communication systems and is distinct from the synthetic training data used during model optimization, as it better captures real-world device characteristics and acoustic variability.

The test dataset is organized into three primary acoustic scenarios: near-end single-talk, far-end single-talk, and double-talk conditions. These scenarios are selected to independently evaluate speech preservation, echo suppression, and robustness under overlapping speech activity. In near-end single-talk conditions, only the local speaker is active, allowing evaluation

Table 5.1: Evaluation scenarios used from the ICASSP AEC Challenge test dataset.

Scenario	**# Files**	**Primary Evaluation Objective**
Near-End Single-Talk	100	Speech preservation and noise suppression
Far-End Single-Talk	100	Echo cancellation performance
Double-Talk	100	Joint echo suppression and speech preservation
Total	**300**	Comprehensive system evaluation

of speech preservation and noise suppression without echo interference. In far-end single-talk conditions, only the remote speaker is active, enabling direct assessment of echo cancellation performance. In double-talk scenarios, both speakers are active simultaneously, representing the most challenging and realistic operating condition for AEC systems, where suppression and preservation must be balanced.

In addition to these acoustic scenarios, the test dataset includes recordings both with and without speaker movement. The movement conditions are designed to simulate changes in the acoustic echo path that can occur in practical hands-free communication systems, such as when a speaker or device changes position relative to the microphone and loudspeaker. These changes introduce a time-varying echo path, requiring the AEC system to continuously adapt its filter coefficients to maintain effective echo cancellation. This provides a particularly challenging condition for linear adaptive AEC systems, as changes in the acoustic environment can result in temporary echo leakage while the adaptive filter converges toward the new echo path.

For each scenario category, 100 test utterances are selected, resulting in a balanced evaluation set across all three conditions. As summarized in Table 5.1, this subset is used to ensure consistent and computationally manageable evaluation while preserving diversity across speakers and acoustic environments.

A balanced distribution across the three acoustic scenarios was intentionally selected to prevent performance metrics from being dominated by any single operating condition. Although conversational speech generally follows a turn-taking structure, overlapping speech remains a common occurrence due to interruptions, listener backchannels, laughter, and

rapid conversational exchanges [91]. Since double-talk represents one of the most challenging scenarios for AEC systems, allocating an equal number of double-talk samples ensures that system robustness under overlapping speech conditions is evaluated with the same statistical significance as the near-end and far-end single-talk cases.

The signal generation process follows a realistic communication pipeline. Far-end speech is passed through a playback chain that may include Acoustic Digital Signal Processing (ADSP) components such as Dynamic Range Compression (DRC), before being emitted by a loudspeaker. The resulting signal propagates through the acoustic environment and is captured by a microphone, where it combines with near-end speech, background noise, and reverberation. This recorded microphone signal serves as the input to the evaluated system. This setup reflects real-world deployment conditions in hands-free communication systems, where both echo and local speech coexist in a reverberant and potentially nonlinearly distorted environment.

Overall, this test configuration provides a controlled yet realistic benchmark for evaluating echo cancellation and noise suppression performance across a range of representative acoustic scenarios. To quantify system performance under these conditions, a set of objective and perceptually motivated evaluation metrics is introduced in the following section.

5.5 Testing Procedure

To ensure reproducibility and consistent evaluation across all test conditions, two complementary evaluation pipelines are used in this work. These correspond to (i) single-file analysis for time-resolved echo behaviour under streaming constraints and (ii) batch evaluation across the ICASSP AEC Challenge test set for perceptual assessment.

The first pipeline operates on individual utterances and is primarily used for analysing temporal echo suppression behaviour using ERLE. In this case, each audio file is processed using a block-wise streaming inference procedure identical to the training configuration,

where intermediate hidden states are carried across consecutive frames. To reflect real-time deployment constraints, inference is performed in a chunk-wise manner, and the reconstructed signal is obtained from overlapping block outputs to ensure temporal alignment of the processed signal.

During processing, ERLE is computed on a per-block basis, producing a time-series representation of echo attenuation over the duration of the utterance. This enables visual inspection of system behaviour under dynamic acoustic conditions, particularly during transitions between far-end single-talk and double-talk segments. The resulting ERLE curves are plotted per file to assess stability and responsiveness of the proposed system.

The second pipeline is designed for large-scale perceptual evaluation using objective quality metrics. In this case, the model is executed over the test set consisting of near-end single-talk, far-end single-talk, and double-talk conditions, with 100 utterances per category. For each file, the enhanced output is generated in a streaming manner consistent with the model inference procedure. For perceptual evaluation, however, the full enhanced STFT is reconstructed prior to applying a single inverse STFT operation. This ensures that no block-wise reconstruction artifacts influence perceptual scoring.

The resulting signals are evaluated using pre-trained AECMOS and DNSMOS models [17, 16]. The scores are stored on a per-file basis and subsequently aggregated to produce scenario-level and overall performance statistics, as described in Section 5.6.

5.6 Evaluation Metrics

The performance of the proposed system is evaluated using three complementary metrics: AECMOS, DNSMOS, and ERLE [17, 16, 1]. Together, these metrics provide a comprehensive assessment of echo cancellation effectiveness, speech quality preservation, and temporal echo suppression behaviour under dynamic acoustic conditions. This section goes into detail on the three metrics.

5.6.1 AECMOS

AECMOS is a perceptual evaluation metric specifically designed to assess the performance of AEC systems in terms of both echo suppression quality and speech preservation. Unlike traditional signal-based metrics, AECMOS is model-driven and trained to predict human perceptual judgments, making it more closely aligned with subjective listening evaluations [17].

The metric is widely used in modern AEC research due to its ability to jointly capture the trade-off between effective echo removal and minimal speech distortion. This is particularly important in real-world systems, where aggressive echo suppression can often degrade near-end speech quality.

AECMOS is composed of two distinct sub-metrics: Echo MOS and Degradation MOS. The Echo MOS component evaluates the effectiveness of echo suppression by measuring the perceptual residual echo level present in the processed signal. Higher Echo MOS values indicate stronger echo cancellation performance. In contrast, the Degradation MOS component measures the degree of distortion introduced to the near-end speech signal during processing. Higher Degradation MOS values correspond to better preservation of speech quality, indicating minimal artificial distortion or speech attenuation.

In this work, AECMOS is calculated over the test set described previously, consisting of 100 single-talk utterances near the end, 100 single-talk utterances far the end, and 100 double-talk utterances. For each scenario, the Echo MOS and Degradation MOS scores are first computed on a per-file basis and then averaged across all 100 utterances to obtain a representative score for that condition. This results in three scenario-level scores: near-end, far-end single-talk, and double-talk, each containing both Echo MOS and Degradation MOS values.

In addition to the per-scenario evaluation, an overall score is computed by averaging the scenario-level results across all three conditions. This provides a single aggregate measure of system performance that captures behaviour across all acoustic environments. This

hierarchical evaluation structure is particularly useful for comparing different model configurations and hyperparameter settings, as it enables both fine-grained analysis of specific acoustic conditions and a global performance comparison across the full test distribution.

While AECMOS effectively captures echo suppression performance and speech degradation, the proposed system also targets noise suppression. Therefore, an additional metric is required to evaluate overall speech quality and background noise reduction, motivating the use of DNSMOS.

5.6.2 DNSMOS

DNSMOS is a non-intrusive, neural network-based metric designed to estimate perceived speech quality [16]. In this context, non-intrusive evaluation methods operate directly on the processed (degraded) signal without requiring access to a clean reference signal. In contrast, intrusive metrics rely on a corresponding ground-truth signal to quantify differences in terms of distortion or error between the enhanced and reference signals.

Unlike traditional intrusive objective metrics, DNSMOS is trained on large-scale subjective listening test data to predict human perceptual judgments of speech quality under noisy and distorted conditions.

The model outputs three perceptual scores: Overall MOS, Speech MOS, and Noise MOS. Each of these components captures a different aspect of perceived audio quality. The Speech MOS Signal (SIG) score evaluates the perceived intelligibility and naturalness of the speech signal, reflecting how clearly and naturally the speech is preserved after processing. The Noise MOS Background (BAK) score measures the perceived level of background noise, indicating how effectively unwanted noise components are suppressed. Finally, the Overall MOS Overall (OVRL) score provides a combined assessment of overall speech quality, balancing both speech preservation and noise reduction.

In this work, DNSMOS is evaluated on the aforementioned test set. For each scenario, the SIG, BAK, and OVRL scores are computed on a per-file basis and then averaged across

all 100 samples to obtain scenario-specific performance metrics.

This results in three sets of averaged scores corresponding to near-end single-talk, far-end single-talk, and double-talk conditions. In addition, an overall score is computed by averaging the three scenario-level results for each of the DNSMOS components. This provides a single aggregated representation of speech quality performance across all acoustic conditions.

This hierarchical averaging strategy enables both fine-grained and global evaluation of system behaviour, mirroring the analysis performed for AECMOS. Scenario-level scores allow detailed analysis of performance under specific acoustic conditions, while the overall averaged scores provide a compact metric for comparing different model configurations and hyperparameter settings across the full evaluation set.

In contrast to the perceptual, reference-free evaluation provided by AECMOS and DNSMOS, the following section introduces ERLE, which provides a signal-level measure of echo attenuation and enables time-resolved analysis of echo suppression performance within individual utterances.

5.6.3 ERLE

ERLE is a signal-level metric used to quantify the amount of echo attenuation achieved by an AEC system [1]. Unlike perceptual metrics such as AECMOS or DNSMOS, ERLE directly measures the reduction in echo power between the input microphone signal and the processed output signal, providing a physically interpretable measure of echo suppression performance based on signal energy reduction.

ERLE is defined as the logarithmic ratio between the power of the microphone signal $d(n)$ and the power of the enhanced output signal $e(n)$ under the assumption of a cancelled echo path:

$$\text{ERLE} = 10 \log_{10} \left(\frac{E[d^2(n)]}{E[e^2(n)]} \right). \tag{5.4}$$

where $E[\cdot]$ denotes the statistical expectation. In practice, these expectations are estimated

using time-averaged signal energies over a finite evaluation window, assuming the signals are ergodic and stationary over the observation interval.

Higher ERLE values indicate stronger echo suppression, as they correspond to a larger reduction in residual echo energy at the output. However, ERLE alone does not distinguish between true echo removal and unintended near-end speech suppression, and must therefore be interpreted alongside perceptual metrics to ensure comprehensive evaluation.

In this work, ERLE is computed on a per-file basis using a block-wise streaming formulation. Each utterance is divided into overlapping frames, and the metric is calculated for each block independently. This results in a time-resolved ERLE trajectory that reflects how echo suppression performance dynamically evolves over the duration of a signal.

This block-wise formulation is particularly useful for analysing system stability across distinct acoustic conditions [1]. In near-end single-talk scenarios, where no far-end signal is present, ERLE naturally approaches 0 dB and remains stable, since there is no echo energy to suppress and the system preserves the near-end speech intact. In far-end single-talk scenarios, ERLE is expected to be high and relatively stable, as the adaptive filters and deep learning architectures can focus entirely on removing linear and nonlinear echo components. In contrast, double-talk scenarios typically exhibit a noticeable drop in measured ERLE. This degradation occurs because the simultaneous presence of active near-end speech power mathematically increases the denominator of the formulation, introducing a structural ambiguity between true echo removal and near-end speech preservation.

As a result, tracking ERLE block-by-block is highly effective for evaluating temporal stability and tracking convergence behavior, especially during rapid transitions between far-end single-talk and double-talk regions. This makes it a valuable diagnostic metric for assessing the real-time robustness of the proposed hybrid system at the utterance level.

Overall, the evaluation framework combines perceptual and signal-level metrics to provide a comprehensive assessment of the proposed AEC system. AECMOS is used to directly quantify echo suppression effectiveness and speech degradation in a perceptually motivated

manner, while DNSMOS evaluates general speech quality in terms of intelligibility, noise suppression, and overall perceptual clarity. In contrast, ERLE provides a physically interpretable, time-resolved measure of echo attenuation, enabling detailed analysis of system behaviour under dynamic acoustic conditions.

Together, these metrics capture complementary aspects of performance, ensuring that improvements are reflected not only in objective echo reduction, but also in preserved speech quality and robustness across varying acoustic scenarios. The following chapter presents the results obtained using this evaluation framework and provides a detailed analysis of system performance across all test conditions.

Chapter 6

Results and Analysis

6.1 Introduction

This chapter presents the results and analysis of the proposed hybrid deep learning-based AEC and NS system, and compares its performance to that of a traditional signal processing pipeline. The evaluation framework, metrics, and test procedures have been defined in Chapter 5.6, and are therefore only briefly referenced here to avoid redundancy.

The analysis is structured around two complementary evaluation approaches. First, a large-scale evaluation is conducted on the ICASSP AEC test set, where performance is quantified using AECMOS and DNSMOS. Scenario-level and overall averages are used to provide a statistically robust comparison across near-end, far-end, and double-talk conditions. Second, a time-resolved analysis based on frame-wise ERLE is used to examine system behaviour within individual utterances, enabling detailed investigation of temporal dynamics during transitions between far-end single-talk, double-talk, and near-end single-talk conditions.

In addition to enhancement performance, practical deployment considerations are evaluated. The proposed model is analyzed in terms of algorithmic latency, refresh rate and RTF, and compared against the deployment requirements of real-world systems. This comparison highlights the trade-offs between enhancement performance and computational efficiency.

Finally, an ablation study is conducted to investigate the impact of key design choices,

including input representations, architectural components, training configurations, and inference strategies such as frame buffering and overlap. This analysis provides insight into the factors that most significantly influence system performance and supports the design decisions adopted in the proposed model.

The remainder of this chapter is organized as follows. Section 6.2 presents the results from the ICASSP test set. Section 6.3 examines performance across different acoustic scenarios. Section 6.4 analyzes temporal behaviour using ERLE. Section 6.5 discusses computational performance and real-time considerations. Section 6.6 presents the ablation study, and Section 6.7 concludes with a discussion of the key findings.

6.2 Objective Results

This section presents the objective evaluation results obtained using AECMOS and DNSMOS across all evaluated acoustic scenarios.

The evaluated systems form a progression, beginning with the unprocessed baseline, followed by a linear AEC system, and incremental additions of AES and NS. The proposed hybrid deep learning-based model is evaluated alongside these systems to provide a comprehensive comparison of performance across increasing system complexity.

6.2.1 AECMOS Results

Table 6.1 presents the AECMOS results across near-end single-talk, far-end single-talk, and double-talk scenarios, along with the overall average.

From Table 6.1, a clear progression in echo suppression performance is observed as additional processing stages are introduced. The baseline and linear AEC systems exhibit lower Echo MOS values in far-end (2.16 and 2.98) and double-talk (2.69 and 3.17) conditions, indicating a clear limitation in their ability to suppress residual and non-linear echo components on their own.

Table 6.1: AECMOS results across different acoustic scenarios

	Near-End		Far-End		Double-Talk		Overall	
Model	Echo	Deg	Echo	Deg	Echo	Deg	Echo	Deg
Baseline	**5.00**	3.90	2.16	**5.00**	2.69	**3.77**	3.28	**4.22**
Linear AEC	**5.00**	**3.91**	2.98	**5.00**	3.17	2.98	3.72	3.96
AEC + AES	**5.00**	3.28	**4.00**	4.99	4.56	1.93	4.52	3.40
AEC + AES + NS	**5.00**	2.60	3.96	4.99	4.53	2.14	4.50	3.24
Proposed Hybrid	**5.00**	3.51	3.96	4.99	**4.64**	2.14	**4.53**	3.55

In the near-end single-talk scenario, where no echo is present, the sequential addition of traditional heuristic blocks causes the Degradation MOS to drop severely, falling from 3.91 for the AEC stage down to 2.60 when both AES and NS are active. This drop reflects the accumulation of musical noise and spectral artifacts generated by independent processing stages. In contrast, the proposed hybrid model mitigates this effect significantly, maintaining a high Degradation MOS of 3.51, which indicates it is better at bypassing unnecessary processing when acoustic distortions are absent.

This architectural advantage is further highlighted in the double-talk scenario. The traditional AEC + AES configuration achieves strong echo attenuation (4.56 Echo MOS) but introduces severe speech distortion, causing the Degradation MOS to fall to a minimum of 1.93. This behavior demonstrates that traditional rule-based residual suppression often aggressively suppresses the entire channel to eliminate echo at the cost of destroying near-end speech intelligibility. The degradation is partially alleviated with the introduction of the NS module. While this improvement may appear counterintuitive, it is likely due to the additional suppression of residual interference by the NS stage, which improves the perceived signal quality measured by AECMOS despite the continued distortion of the near-end speech.

The proposed hybrid model on the other hand successfully reduces this trade-off. It achieves both the highest double-talk Echo MOS (4.64) and matches the highest Degradation MOS (2.14) of the full pipeline. By utilizing a continuous CRM target, the deep learning post-filter is able to jointly model the speech and echo components in the complex spectral domain, decoupling residual echo cancellation from direct signal attenuation. Therefore, the proposed

architecture provides better non-linear suppression while minimizing the destructive over suppression typical of cascaded pipelines.

Finally, in far-end single-talk scenarios, the proposed model matches the high suppression capabilities of the fully cascaded baseline (3.96 vs 3.96), while maintaining an optimal Degradation MOS of 4.99. Cumulatively, these behaviors allow the proposed model to achieve the highest overall Echo MOS (4.53) and Overall Degradation MOS (3.55), validating the effectiveness of the hybrid approach across balanced conversational conditions.

6.2.2 DNSMOS Results

Table 6.2 presents the DNSMOS results across all scenarios, including Speech MOS (SIG), Noise MOS (BAK), and Overall MOS (OVRL).

Table 6.2: DNSMOS results across different acoustic scenarios

	Near-End			Far-End			Double-Talk			Overall		
Model	SIG	BAK	OVRL	SIG	BAK	OVRL	SIG	BAK	OVRL	SIG	BAK	OVRL
Baseline	**3.78**	3.51	3.84	**2.96**	**2.62**	2.81	**3.08**	**2.92**	2.77	**3.27**	**3.02**	3.14
Linear AEC	**3.78**	**3.52**	3.85	2.66	1.92	2.18	2.82	2.36	2.29	3.09	2.60	2.77
AEC + AES	3.50	3.28	3.85	2.23	1.49	2.48	2.49	2.07	2.65	2.74	2.28	2.99
AEC + AES + NS	3.18	2.78	3.82	2.22	1.91	3.61	2.41	2.16	3.63	2.60	2.28	3.69
Proposed Hybrid	3.73	3.29	**4.05**	2.29	1.76	**3.79**	2.67	2.32	**3.85**	2.90	2.46	**3.90**

The DNSMOS results indicate that the proposed hybrid model achieves the highest OVRL across all evaluated scenarios. It is important to note that the OVRL score is a separate perceptual estimate of overall speech quality and is not calculated as a simple average of the SIG and BAK scores. Instead, it provides an overall assessment of the perceived quality of the processed speech, taking into account both the quality and clarity of the near-end speech and the quality of the background or interfering signal. As a result, improvements in OVRL do not necessarily correspond to improvements in either SIG or BAK individually. The higher OVRL scores achieved by the proposed hybrid model therefore suggest that it provides a more favorable overall balance between speech preservation and interference suppression. By replacing the cascaded modules with a single learned framework, the hybrid system can jointly account for these competing objectives rather than optimizing each

processing stage independently.

Examining the near-end single-talk conditions reveals a significant weakness in the traditional cascaded pipeline. When the NS module is added to the AEC + AES system, the Noise MOS (BAK) score drops from 3.28 to 2.78. This indicates that traditional noise suppression can actually worsen the perceived background quality when the acoustic environment is already relatively clean. Because conventional NS relies on noise estimates derived from the input signal, applying it sequentially after earlier processing stages can result in inaccurate noise estimates. The system may consequently suppress signal components that do not require attenuation, introducing perceptual artifacts that negatively affect the resulting speech quality. The proposed hybrid model avoids this additional cascaded processing stage, achieving a BAK score of 3.29 while also obtaining the highest near-end OVRL score of 4.05.

Furthermore, the hybrid model outperforms the complete traditional pipeline (AEC + AES + NS) in almost every metric across all scenarios, with one notable exception: BAK performance during far-end single-talk (1.76 for the hybrid vs. 1.91 for the traditional pipeline). This specific result highlights the difference in how the two systems handle suppression. In far-end single-talk, there is no near-end speaker to protect, allowing the traditional NS to ruthlessly mute the microphone line. While this aggressive suppression achieves a slightly higher BAK score, it is a rigid approach that fails when a real conversation occurs.

The hybrid model, conversely, takes a slightly more conservative approach to background suppression to ensure it does not accidentally clip speech. The payoff for this design choice is clear in the double-talk scenario. While the traditional pipeline struggles with overlapping audio, the hybrid model achieves significantly higher scores across the board during double-talk (SIG 2.67, BAK 2.32, OVRL 3.85).

Overall, performance degradation is most apparent across all systems in double-talk scenarios, which remain the most challenging condition due to the simultaneous presence of near-end speech and residual echo. Despite this difficulty, the hybrid model demonstrates superior robustness. It demonstrates that jointly predicting a complex mask for both echo

and noise is a more effective strategy than chaining separate rule-based estimators together.

6.2.3 Overall Trends

Across all evaluated scenarios, a consistent set of trends emerges when comparing system complexity and performance. The linear AEC systems provide modest improvements over the unprocessed condition, primarily reflected in improved echo-related and overall perceptual scores. However, their performance remains limited in more challenging acoustic conditions, particularly during double-talk, where residual echo and overlapping speech significantly degrade performance.

As additional processing stages are introduced in the form of AES and NS, improvements in interference suppression become more pronounced in specific scenarios. However, these gains are accompanied by a noticeable reduction in speech quality consistency, particularly in near-end single-talk conditions. This indicates that cascaded processing pipelines tend to over-apply suppression even when it is not required, leading to unnecessary distortion and reduced perceptual quality.

In contrast, the proposed hybrid model demonstrates a more balanced performance profile across all metrics and scenarios. It consistently maintains high speech quality while achieving strong echo and noise suppression, particularly in double-talk conditions where traditional systems typically struggle. Importantly, it mitigates the trade-offs observed in the cascaded pipeline, where improvements in one metric often come at the expense of another.

To further illustrate these trends, Figure 6.1 presents an aggregated comparison of the key overall metrics across all systems, including Echo MOS, Degradation MOS, and DNSMOS OVRL.

Overall, these results suggest that jointly modelling echo and noise suppression within a single learned representation leads to improved robustness across varying acoustic conditions. These results are fleshed out and explored further in the next section where individual components are covered in more detail.

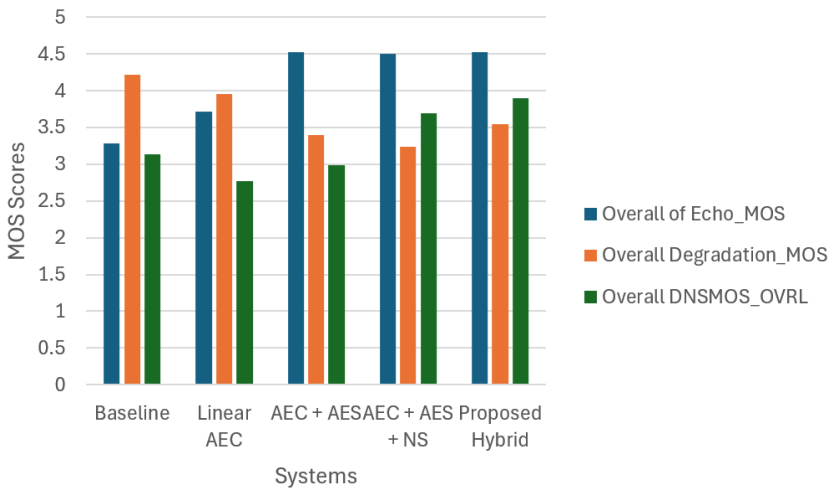


Figure 6.1: Overall performance comparison across all evaluated systems. The proposed hybrid model achieves the best balance between echo suppression, speech preservation, and perceptual quality, outperforming both the traditional cascaded pipeline and intermediate baselines.

6.3 Scenario-Based Performance Analysis

This section provides a detailed analysis of system performance across different acoustic scenarios, namely near-end single-talk, far-end single-talk, and double-talk conditions. While Section 6.2 presents aggregated quantitative results, the focus here is on interpreting how each system behaves under specific acoustic conditions and how architectural design choices influence performance.

The analysis is structured in two parts. First, the progression of the traditional signal processing pipeline is examined, starting from the baseline system and incrementally incorporating linear AEC, AES, and NS. This allows for a step-by-step understanding of how each component contributes to overall performance.

Second, the proposed hybrid deep learning model is analyzed independently to highlight its behaviour under the same acoustic conditions. A comparative discussion is then provided to contrast its performance characteristics with those of the traditional pipeline, particularly in terms of robustness, adaptability, and perceptual quality.

6.3.1 Traditional Signal Processing Pipeline Analysis

This subsection analyzes the behaviour of the traditional signal processing pipeline across progressively more complex system configurations. The focus is on understanding how each additional processing stage addresses specific limitations of the preceding system, particularly with respect to echo cancellation, non-linear distortions, and double-talk robustness.

Baseline System (No Processing)

The baseline system corresponds to the raw microphone signal without any echo cancellation or enhancement processing. As such, it represents the unprocessed acoustic mixture and serves as the reference point for all subsequent system configurations. The AECMOS and DNSMOS results for the baseline system across the evaluated acoustic scenarios are shown in Figure 6.2.

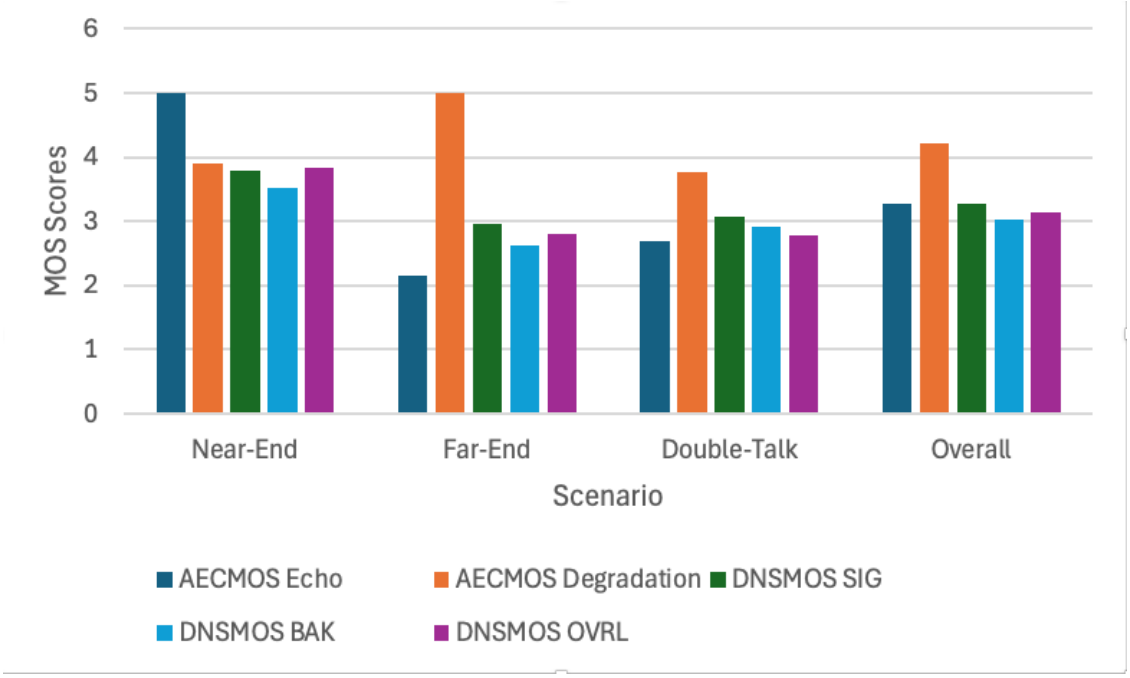


Figure 6.2: AECMOS and DNSMOS results for the baseline system across near-end single-talk, far-end single-talk, double-talk, and overall conditions.

In terms of AECMOS, the expected behaviour is observed in the Echo MOS component. As shown in Figure 6.2, far-end single-talk conditions produce the lowest Echo MOS value

of 2.16, reflecting the unattenuated presence of echo in the signal. In double-talk conditions, the Echo MOS increases slightly to 2.69. This increase does not indicate a reduction in echo energy, but rather reflects the perceptual influence of the near-end speech, which can partially mask the echo. Near-end single-talk conditions naturally achieve a perfect Echo MOS of 5.00, as no far-end speech is active to generate an echo.

The Degradation MOS component follows a different trend. Because no algorithmic processing is applied, the baseline does not introduce processing-related distortion. Therefore, it achieves the highest Degradation MOS values across the evaluated systems in near-end single-talk (3.90), far-end single-talk (5.00), and double-talk (3.77) conditions. These results demonstrate the advantage of avoiding additional processing artifacts, while also highlighting the trade-off associated with leaving the acoustic interference unprocessed.

The DNSMOS results shown in Figure 6.2 further illustrate this trade-off. The baseline system exhibits relatively strong perceptual scores compared with several of the intermediate processing configurations. Most notably, the double-talk Overall MOS (OVRL) is 2.77, compared with 2.29 for the linear AEC and 2.65 for the AEC + AES configurations. This demonstrates that the introduction of additional processing does not necessarily improve perceived speech quality when the processing stages introduce distortions or other perceptual artifacts.

The Speech MOS (SIG) component remains relatively strong in near-end single-talk conditions, with a score of 3.78, since the speech signal is not subjected to algorithmic processing. The score decreases to 2.96 in far-end single-talk and 3.08 in double-talk, where the near-end speech is affected by the presence of acoustic echo and overlapping signals. Similarly, the Noise MOS (BAK) score of 3.51 in near-end single-talk provides a reference for the background conditions present in the dataset before active noise suppression is applied.

Overall, the baseline system illustrates the fundamental trade-off between interference suppression and signal preservation. By applying no processing, the system avoids algorithmic degradation and consequently maintains relatively strong Degradation and Speech MOS

scores. However, the absence of echo and noise suppression results in substantially poorer performance in conditions where acoustic interference is present. The baseline therefore provides an important reference against which the benefits and potential perceptual costs of the subsequent processing stages can be evaluated.

Linear AEC (PBFDAF)

The linear AEC stage introduces the PBFDAF to explicitly estimate and subtract the linear acoustic echo path. Compared to the unprocessed baseline, this configuration represents the system's first active intervention to isolate the far-end echo from the near-end mic signal. The resulting AECMOS and DNSMOS scores across the evaluated acoustic scenarios are shown in Figure 6.3.

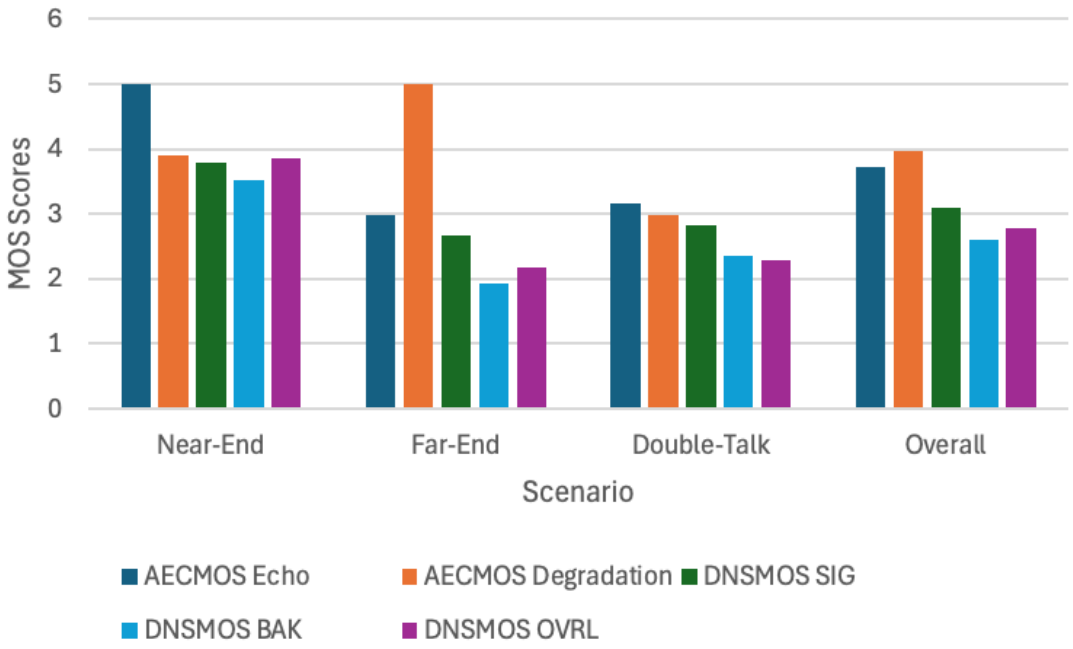


Figure 6.3: AECMOS and DNSMOS results for the linear AEC system across near-end single-talk, far-end single-talk, double-talk, and overall conditions.

In terms of AECMOS, the system demonstrates clear progress in raw echo reduction. As shown in Figure 6.3, in far-end single-talk, the Echo MOS rises from 2.16 to 2.98, and in double-talk conditions, it moves from 2.69 to 3.17. These increases confirm that the adaptive

filter successfully models and attenuates a noticeable portion of the primary linear echo path. Near-end single-talk remains completely stable (5.00 Echo MOS and 3.91 Degradation MOS), proving that the filter remains passive and does not alter the signal when far-end excitation is absent.

However, this gain in echo suppression triggers a major performance penalty in conversational scenarios. During double-talk, the Degradation MOS drops considerably from the baseline score of 3.77 down to 2.98. This drop exposes the mathematical vulnerability of linear adaptive filters.

The adaptation loop relies on an error signal calculated as the difference between the microphone input and the filter's echo estimate. In single-talk, this error consists almost entirely of residual echo, which provides clean feedback for updating the filter coefficients. During double-talk, however, the near-end speaker's voice enters the microphone. To the adaptive filter, this desired near-end speech acts as a massive, uncorrelated noise spike that corrupts the error signal.

To combat this, the proprietary linear AEC stage incorporates a dedicated DTD module designed to freeze adaptation when near-end speech is detected. In practice, however, this detection is inherently imperfect. Real-world decision thresholds struggle with detection latency, meaning that the initial onset of near-end speech often bleeds through and corrupts the filter coefficients before the loop can freeze, resulting in immediate speech distortion.

Furthermore, even if this proprietary DTD were theoretically flawless and achieved instantaneous detection, a fundamental architectural limitation remains: freezing the adaptation loop leaves the filter blind. By suspending updates to protect the coefficients, the system becomes unable to track real-time variations in the acoustic echo path caused by things like subtle physical movements or changing enclosure dynamics. This control loop struggle creates audible mistracking artifacts and clipping, which directly explains the poor 2.98 Degradation MOS score.

This behavioural flaw is heavily penalized in the DNSMOS results shown in Figure 6.3.

While near-end metrics hold steady, the far-end and double-talk scores degrade across the board. In far-end single-talk, Speech MOS falls from 2.96 to 2.66, and Noise MOS plunges from 2.62 to 1.92, showing that the linear filter introduces spectral artifacts that the model interprets as poor background quality.

This can also be seen in the double-talk Overall MOS, which drops from 2.77 to 2.29. This confirms that the perceptual damage caused by filter mistracking outweighs some of the benefits of its echo reduction. Ultimately, these results demonstrate that a purely linear adaptive filter is not enough for full-duplex communication on its own, creating a need for non-linear post-filtering stages.

AEC + AES

The addition of an AES stage introduces a non-linear post-processing component applied to the residual error signal. While the linear adaptive filter handles the primary echo path, the AES is designed to suppress what the filter misses, including residual linear echo and non-linearities caused by loudspeaker distortion, physical enclosure vibrations, and clipping. The resulting AECMOS and DNSMOS scores across the evaluated acoustic scenarios are shown in Figure 6.4.

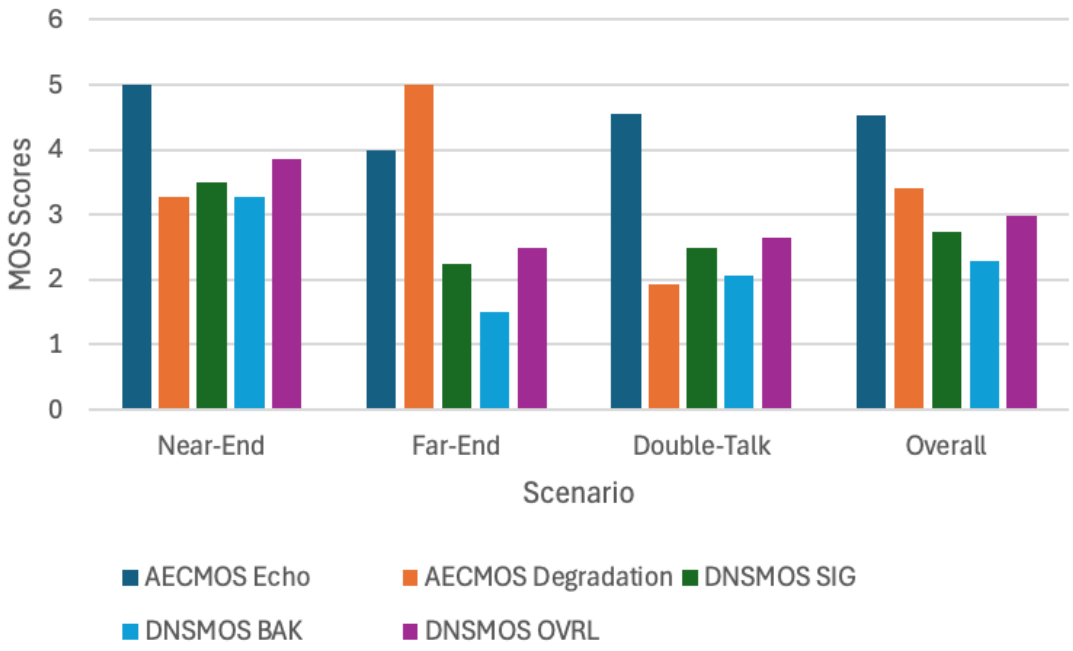


Figure 6.4: AECMOS and DNSMOS results for the AEC + AES system across near-end single-talk, far-end single-talk, double-talk, and overall conditions.

In terms of AECMOS, the integration of AES yields an immediate improvement in echo attenuation. As shown in Figure 6.4, in far-end single-talk conditions, Echo MOS rises from 2.98 to 4.00, and in double-talk, it goes from 3.17 to 4.56. These figures show that the non-linear suppression logic is successful at its primary task: identifying and removing residual echo energy that the linear front-end could not model.

However, this aggressive echo removal comes at a cost to speech quality. During double-talk, the Degradation MOS drops from 2.98 down to 1.93. Furthermore, a highly revealing drop occurs even in near-end single-talk, where Degradation MOS slips from 3.91 to 3.28. This drop in a clean, echo-free scenario indicates that the AES module's statistical estimators can introduce unnecessary attenuation artifacts and spectral roughness even when the local speaker is talking in isolation.

The reason for this degradation during double-talk stems from how traditional rule-based AES modules calculate their attenuation curves. These algorithms typically compute a real-time suppression gain based on the ratio of estimated residual echo power to total

microphone power. During double-talk, near-end speech and residual echo inevitably overlap within the same time-frequency bins. Because a traditional cascaded system lacks the source-separation capability to distinguish these mixed signals, the gain filter can apply attenuation across shared frequency components. The result is voice clipping, audible dropouts, and a fractured speech structure, which is consistent with the poor 1.93 Degradation MOS score.

This perceptual damage is mirrored in the DNSMOS evaluations shown in Figure 6.4. During double-talk, Speech MOS SIG falls from 2.82 to 2.49, and Noise MOS BAK drops from 2.36 to 2.07, indicating that the listening model perceives the processed speech as more distorted and unnatural.

Interestingly, the far-end single-talk Overall MOS shows a slight improvement, rising from 2.18 to 2.48. This occurs because, in the absence of a local speaker, the metric heavily rewards the elimination of echo, finding a line with reduced residual echo and minor filtering artifacts more tolerable than a line containing substantial residual echo. This trend is also seen in the double-talk Overall MOS, which moves up from 2.29 to 2.65.

Ultimately, the AEC + AES configuration demonstrates the cost of traditional cascaded pipelines. Maximizing echo suppression through rigid, rule-based gain filtering forces a trade-off in speech preservation, highlighting the potential benefits of a unified optimization target.

AEC + AES + NS

The final stage of the traditional processing pipeline integrates a NS module following the linear AEC and AES blocks. This configuration represents the classical industry approach to jointly tackling echo, non-linear residual echo components, and ambient background noise within a cascaded, multi-stage architecture. The resulting AECMOS and DNSMOS scores across the evaluated acoustic scenarios are shown in Figure 6.5.

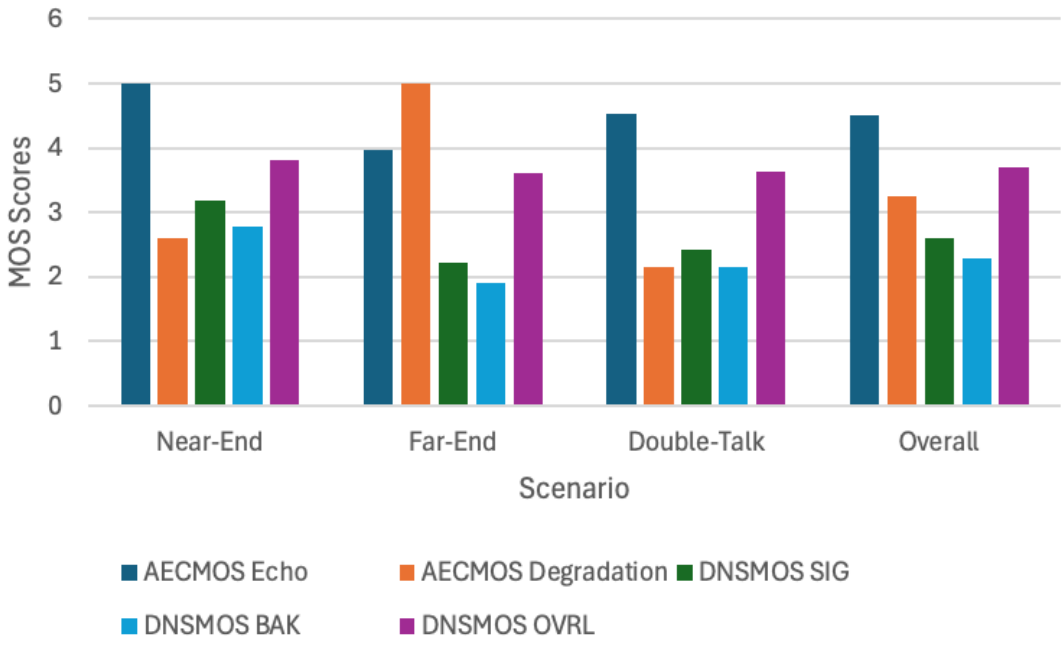


Figure 6.5: AECMOS and DNSMOS results for the AEC + AES + NS system across near-end single-talk, far-end single-talk, double-talk, and overall conditions.

From an AECMOS perspective, adding the NS stage does not fundamentally alter the raw echo suppression performance, which is expected since echo attenuation is primarily handled by the preceding AEC and AES blocks. As shown in Figure 6.5, the Echo MOS remains largely stationary, experiencing only negligible variations (shifting from 4.56 to 4.53 in double-talk and from 4.00 to 3.96 in far-end single-talk).

The Degradation MOS scores reveal an interesting conflict here. During double-talk, the score actually improves slightly, moving from 1.93 to 2.14. This happens because the NS block acts as a sort of spectral smoother, rounding off some of the harsh musical noise left behind by the aggressive AES gating. However, looking at the near-end single-talk scenario, where there is no echo to begin with, adding the NS block penalizes the system, dragging the score down from 3.28 to 2.60. This shows the downside of chaining independent processing stages: it creates a cumulative distortion that progressively degrades clean speech.

This cumulative damage is shown in more detail in the DNSMOS metrics in Figure 6.5, where highly scenario-dependent behaviour and tracking failures emerge.

Most notably, an anomaly occurs in the near-end single-talk condition: the introduction of the noise suppressor causes the Background Noise MOS BAK to paradoxically degrade, dropping from 3.28 down to 2.78. An active NS block should improve or at least maintain the background noise score in an already clean near-end environment, but the opposite happens here. This failure likely occurs because traditional rule-based NS algorithms rely on tracking the signal's minimum statistics to build an accurate noise floor estimate. Because the preceding AEC and AES stages have already processed the signal and introduced non-linear spectral fluctuations, the NS noise estimator becomes confused. It misinterprets these residual algorithmic artifacts as non-stationary background noise, leading to aggressive, erratic filtering that actively corrupts the noise floor.

In contrast, when evaluating scenarios heavily polluted by external interference, the system shows a clear perceptual benefit in the Overall MOS (OVRL) scores. In far-end single-talk, the OVRL score jumps from 2.48 to 3.61, and in double-talk, it climbs from 2.65 to 3.63. These sharp increases occur because the perceptual evaluation model rewards the raw attenuation of background noise and residual echo, finding a heavily muted noise floor more tolerable than one filled with constant ambient noise. However, this overall improvement comes at the expense of speech clarity, as evidenced by the Speech MOS SIG dropping to its lowest points (2.22 in far-end and 2.41 in double-talk).

Ultimately, the fully cascaded AEC + AES + NS pipeline exposes the fundamental limit of traditional signal processing architectures. Because each block operates blindly with little to no awareness of the downstream or upstream modules, their mathematical assumptions routinely clash. Attempting to tune this chain becomes a perpetual compromise: fixing background noise in double-talk directly causes the system to distort clean speech during near-end single-talk. This systemic behaviour demonstrates the necessity of moving away from chained, isolated estimators and instead adopting a unified neural architecture capable of optimizing speech preservation, echo removal, and noise suppression simultaneously.

Proposed Hybrid Deep Learning Model

The proposed hybrid deep learning model provides a unified, end-to-end alternative to the traditional cascaded pipeline by handling echo suppression and noise reduction simultaneously. Instead of treating these tasks as isolated problems, the neural post-filter operates directly in the complex STFT domain. By estimating a CRM, the network jointly updates magnitude and phase information. This joint modeling is a major advantage during double-talk conditions, where preserving the phase structure of the overlapping signals is important for preventing near-end speech distortion. The resulting AECMOS and DNSMOS scores across the evaluated acoustic scenarios are shown in Figure 6.6.

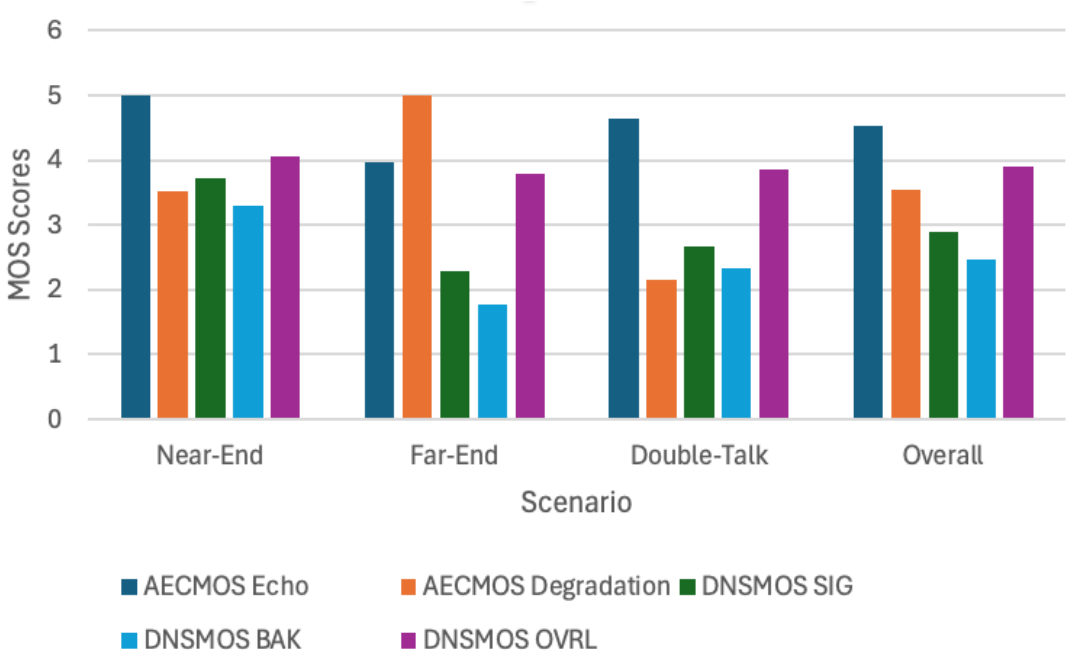


Figure 6.6: AECMOS and DNSMOS results for the proposed hybrid deep learning model across near-end single-talk, far-end single-talk, double-talk, and overall conditions.

From an AECMOS perspective, the proposed model achieves the highest Echo MOS in double-talk conditions (4.64), outperforming every traditional configuration. As shown in Figure 6.6, this indicates that the deep learning model suppresses non-linear residual echo more effectively than the classical configurations. It also achieves this high echo attenuation while matching the best traditional system Degradation MOS score (2.14). This

demonstrates that the hybrid approach successfully mitigates the trade-off seen in traditional systems, reducing residual echo without resorting to the aggressive, switch-like gating behaviour that can clip the local speaker's voice.

The advantage of a unified architecture is even more apparent in the near-end single-talk results. In this scenario, the proposed model achieves a Degradation MOS of 3.51, representing a major improvement compared to the 2.60 score of the full AEC + AES + NS pipeline. Because the deep network learns a direct mapping from the corrupted input features to the target clean speech, it can learn when an environment is free of echo. As a result, it avoids the cumulative degradation of the cascaded baseline, where misaligned noise estimators can continuously introduce artifacts into clean audio.

From a perceptual standpoint, the DNSMOS evaluations shown in Figure 6.6 further highlight the robustness of the hybrid approach. During double-talk, the model secures the highest Overall MOS (3.85) alongside clear improvements in both Speech MOS (2.67) and Noise MOS (2.32). By framing the problem as a single source-separation task, the network effectively isolates and preserves the target near-end speech structure even when it shares time-frequency bins with loud residual echo.

The only metric where the traditional pipeline scores higher is the background noise category BAK during far-end single-talk, where the hybrid model scores 1.76 compared to the traditional pipeline's 1.91. However, this minor drop represents a design trade-off. While the traditional system achieves a higher BAK score by completely muting the line when no local speech is detected, that rigid approach can fail when a conversation begins. The hybrid model maintains a more conservative suppression margin to protect speech onsets from clipping. This results in an improvement in the far-end Overall MOS scores, where the hybrid model outperforms the traditional pipeline (3.79 vs. 3.61).

Ultimately, the proposed model achieves the highest Overall DNSMOS score (3.90) across all acoustic conditions. These results support the core hypothesis of this work: optimizing a single neural model to jointly address residual echo and noise is a promising alternative

to chaining independent, rule-based signal processing blocks together. By replacing rigid thresholds with a flexible, learned representation, the hybrid architecture delivers a balanced full-duplex communication experience across both clean single-talk environments and challenging double-talk scenarios.

6.4 Temporal Behaviour Analysis (Single-File ERLE)

This section presents a time-resolved analysis of echo suppression performance using frame-wise ERLE measured over an 11-second test utterance. The analysis uses the file `D0AaY06jVka vntbg01IFzQ doubletalk with movement` from the Interspeech 2021 test set of the ICASSP AEC Challenge [12]. The recording was selected because it contains distinct single-talk and double-talk regions, as well as speaker movement, providing a challenging and representative scenario for evaluating echo cancellation performance under varying acoustic conditions.

Unlike the scenario-level results presented in Section 6.2, this analysis focuses on temporal dynamics, specifically how quickly each system adapts to changing acoustic conditions and how stable the echo suppression remains over time.

To facilitate interpretation of the time-resolved plots, the different acoustic conditions are highlighted using transparent coloured regions. Far-end single-talk regions are indicated in green, double-talk regions are indicated in orange, and near-end single-talk regions are indicated in magenta. These highlighted regions correspond to the temporal locations of each scenario within the test utterance and provide a visual reference for relating changes in ERLE to the active acoustic condition.

6.4.1 Linear AEC Baseline

Figure 6.7 shows the ERLE behaviour of linear AEC using proprietary PBFDAF algorithm.

The linear AEC system exhibits a pronounced adaptation period at the beginning of the recording. During the initial far-end single-talk segment, which occupies approximately

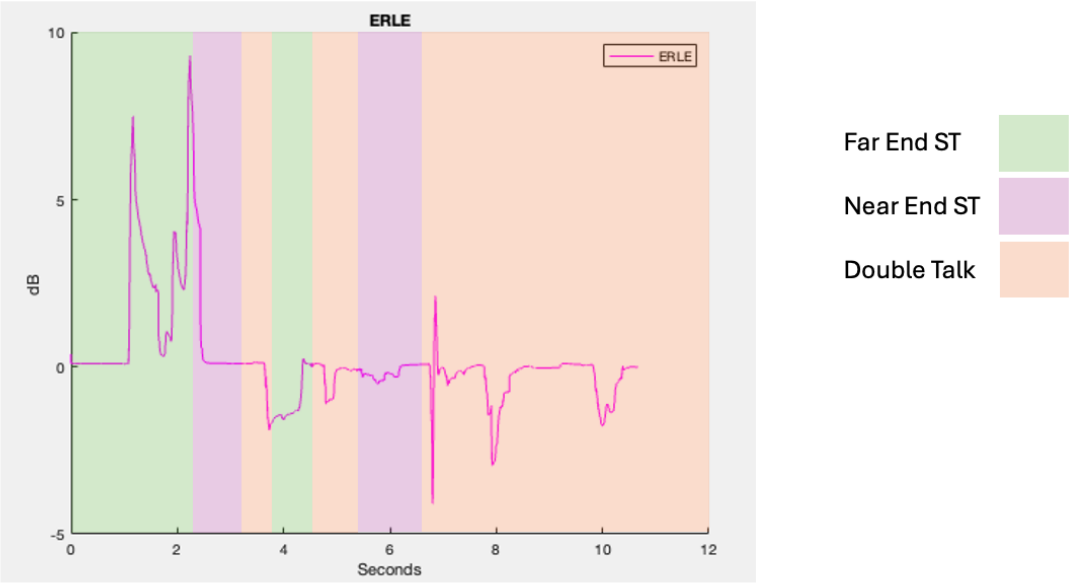


Figure 6.7: ERLE over time for linear AEC using proprietary PBFDAF Algorithm

the first two seconds of the test file, the measured ERLE remains close to 0 dB for most frames, with only intermittent positive peaks. These brief increases in ERLE suggest that the adaptive filter is occasionally producing a partial estimate of the echo path; however, the suppression performance remains highly unstable and does not achieve sustained echo reduction before the onset of double-talk. This behaviour is likely caused by speaker movement within the recording. As the speaker moves, the acoustic echo path continuously changes, forcing the linear filter to repeatedly adapt and re-converge to the evolving environment.

Following the start of the double-talk segment at approximately 2 s, the ERLE decreases further and remains close to 0 dB, with several periods exhibiting negative values. This behaviour indicates that the linear AEC provides little effective echo suppression under these conditions. The poor performance can be attributed to the interaction between the time-varying acoustic environment and the DTD. As double-talk is detected, adaptation of the linear filter is suspended to prevent divergence. However, because the filter has not yet obtained an accurate estimate of the changing echo path, pausing adaptation limits its

ability to track the acoustic variations, resulting in minimal echo attenuation throughout the remainder of the recording.

6.4.2 Linear AEC + AES

Figure 6.8 shows the linear AEC system augmented with AES system.

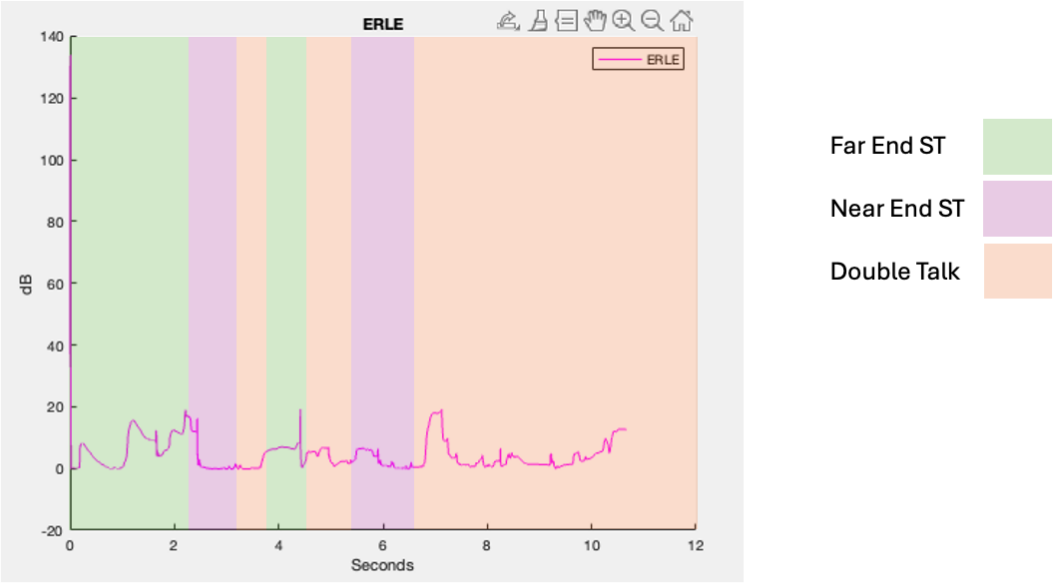


Figure 6.8: ERLE over time for linear AEC with AES

The addition of the AES stage significantly improves echo suppression performance compared to the linear AEC alone. During the initial far-end single-talk segment, the ERLE curve is substantially more stable, remaining predominantly within the 5–15 dB range despite occasional gradual drops. This behaviour indicates that the AES stage is able to provide consistent residual echo attenuation even when the linear adaptive filter has not fully converged. A large transient spike is observed at the very beginning of the recording, where the measured ERLE briefly exceeds 100 dB. This spike is likely caused by very low residual signal energy during the initial frames, which can artificially inflate the ERLE calculation and is therefore not representative of steady-state system performance.

Informal subjective listening confirms the objective results observed during the single-talk region. The dominant echo component is effectively removed, leaving primarily background noise with little audible far-end speech leakage. This represents a substantial improvement over the linear AEC alone, which was unable to achieve stable echo suppression under the same conditions.

Following the onset of double-talk, the ERLE exhibits greater variability, with periods of both increased and decreased suppression. This behaviour is expected, as the presence of near-end speech increases the energy of the microphone signal and generally reduces the measured ERLE. Despite these fluctuations, the ERLE remains predominantly above 0 dB and typically lies within the 0–15 dB range throughout the double-talk segment. This demonstrates that the AES stage successfully mitigates the performance degradation observed with the linear AEC alone and continues to provide meaningful echo attenuation even under challenging double-talk conditions.

From a perceptual perspective, the output during double-talk contains the desired near-end speech with only limited residual echo. However, occasional speech distortion is audible, particularly during more challenging portions of the recording. This distortion is likely caused by inaccuracies in the residual echo estimate supplied by the linear AEC. When the estimated echo power exceeds the true residual echo level, the suppression stage may apply excessive attenuation, resulting in temporary over-suppression of the desired near-end speech. Therefore, although the AES stage substantially improves echo reduction, its performance remains dependent on the quality of the linear echo estimate provided to the suppression algorithm.

6.4.3 Linear AEC + AES + NS

Figure 6.9 shows the full traditional pipeline including noise suppression.

The addition of the NS stage further increases the measured ERLE relative to the AEC+AES configuration. During the initial far-end single-talk segment, the ERLE remains

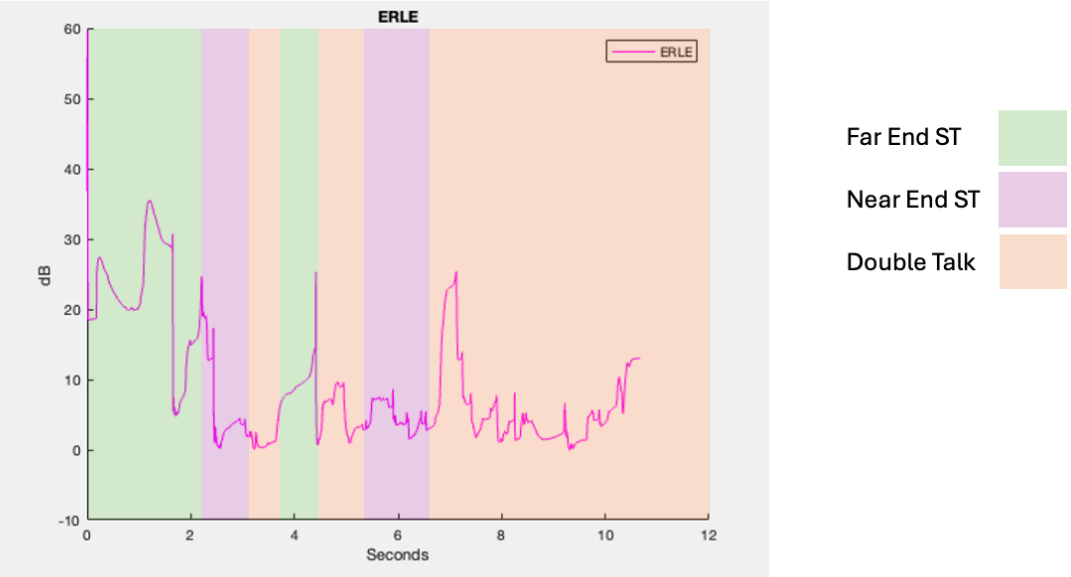


Figure 6.9: ERLE over time for linear AEC + AES + NS

predominantly within the 20–35 dB range, with only occasional decreases to approximately 5 dB. Unlike the linear AEC system, the ERLE does not collapse to 0 dB, indicating that the combined suppression stages maintain consistent attenuation throughout the single-talk period.

During the double-talk segment, the ERLE exhibits the expected fluctuations associated with changing near-end speech activity. While brief periods approach 0 dB, the majority of values remain well above this level, with peaks reaching 20–30 dB and lower values generally remaining within the 5–10 dB range. Overall, the ERLE is consistently higher than that observed for the AEC+AES configuration. This improvement is expected, as the noise suppression stage removes additional signal energy beyond residual echo, reducing the residual power measured at the output and consequently increasing the calculated ERLE.

Although the objective performance appears highly favourable, informal subjective listening reveals a more nuanced result. The addition of the NS stage substantially reduces the background noise that remained after echo suppression, producing a noticeably cleaner output signal. However, this improvement is accompanied by increased speech distortion,

particularly during challenging portions of the recording. The distortion is likely caused by inaccuracies in the suppression gains applied by both the residual echo suppressor and the noise suppression stage. Because the recording contains speaker movement and therefore a continuously changing acoustic environment, the underlying echo and noise estimates may not accurately track the true signal characteristics at all times.

Furthermore, the cascaded structure of the traditional enhancement pipeline makes the system particularly susceptible to error propagation. The NS stage operates on the output of the preceding AEC and AES stages and therefore inherits any inaccuracies present in their residual signal estimates. If the linear AEC fails to accurately model the time-varying echo path, the resulting residual echo estimate supplied to the AES stage becomes less reliable. Any over- or under-suppression introduced by the AES stage is subsequently passed to the NS stage, which may further amplify these errors through additional attenuation. This results in estimation errors that accumulate throughout the processing chain, increasing the likelihood that portions of the desired near-end speech will be misclassified as interference and suppressed.

As a result, portions of the desired speech can be mistakenly attenuated, leading to over-suppression and perceptible speech degradation. What this translates to is, while the AEC+AES+NS system achieves the highest objective ERLE performance among the traditional approaches, the increased suppression comes at the cost of reduced speech quality. This observation highlights a fundamental limitation of cascaded enhancement systems: improvements in traditional interference suppression signal-based metrics like ERLE do not necessarily translate to improved perceptual quality when estimation errors propagate through multiple sequential processing stages.

6.4.4 Proposed Hybrid Deep Learning Model

It should be noted that the appearance of the ERLE plot differs slightly from the previous figures because the proposed deep learning system was evaluated and visualized using

a Python-based framework, whereas the traditional systems were evaluated using a proprietary MATLAB implementation. Therefore, differences in line colours, plotting styles, and figure formatting may be observed; however, the underlying ERLE values and calculation methodology remain unchanged.

Figure 6.10 presents the ERLE behaviour of the proposed hybrid deep learning model.

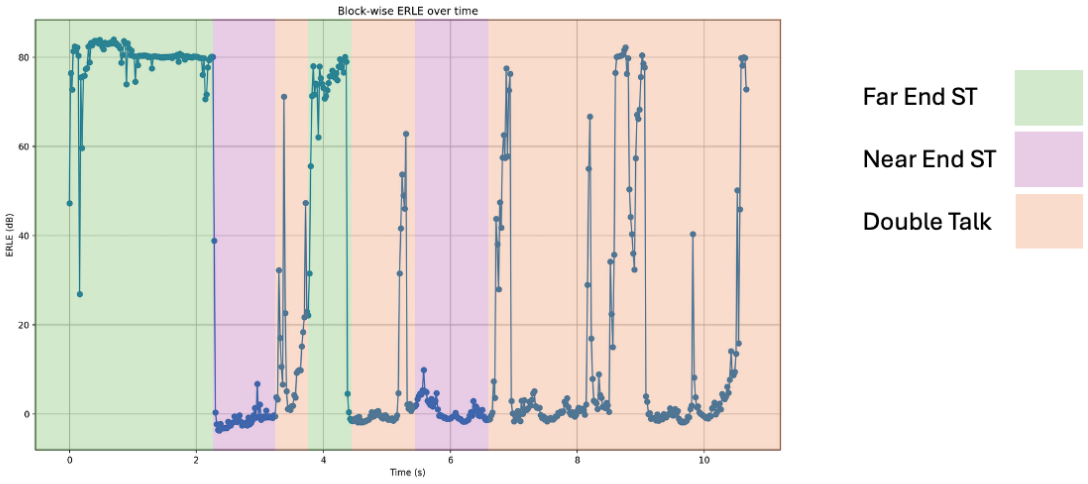


Figure 6.10: ERLE over time for proposed hybrid deep learning model

The proposed hybrid AEC system, consisting of a linear AEC followed by the deep learning post-filter, exhibits substantially stronger performance than any of the traditional cascaded approaches. During the initial far-end single-talk segment, the measured ERLE remains consistently close to 80 dB, with only minor fluctuations. This represents a significant improvement over the linear AEC, AEC+AES, and AEC+AES+NS configurations, demonstrating the ability of the deep learning post-filter to effectively suppress residual echo even when the linear front-end has not produced an accurate echo estimate.

Following the onset of double-talk, the ERLE exhibits a characteristic pattern in which brief reductions toward 0 dB are followed by an almost immediate recovery to approximately 80 dB. These reductions coincide with periods of near-end speech activity, where the presence of the desired speech signal naturally affects the ERLE measurement. More importantly, the

rapid recovery observed throughout the remainder of the recording demonstrates the ability of the proposed model to adapt effectively to changing acoustic conditions and transitions between far-end single-talk and double-talk. Unlike the traditional approaches, which exhibited prolonged performance degradation once adaptation errors occurred, the proposed model maintains consistently high suppression performance across the entire recording.

A particularly notable result is that this level of performance is achieved despite the relatively poor quality of the linear AEC output and echo estimate supplied to the network. The PBFDAF front-end struggled to track the time-varying echo path introduced by speaker movement, resulting in weak standalone performance. Nevertheless, the deep learning post-filter is able to compensate for these inaccuracies and continue providing strong echo suppression. This behaviour highlights one of the key advantages of the proposed architecture: the model is not solely dependent on the accuracy of the linear echo estimate and can learn to suppress residual interference directly from the available acoustic cues.

Informal subjective listening results align closely with the objective measurements. Throughout the recording, both the residual echo and background noise are effectively removed, while the desired near-end speech remains clear and intelligible. Unlike the traditional cascaded pipeline, which exhibited increasing speech distortion as additional suppression stages were introduced, the proposed model maintains speech quality while simultaneously achieving superior echo attenuation. This demonstrates the benefit of performing residual echo suppression and noise reduction within a unified deep learning framework, particularly in challenging scenarios involving speaker movement and frequent double-talk activity.

6.4.5 Overall Temporal Trends

Overall, the comparative analysis across the four configurations highlights a clear progression in echo suppression capability and system behaviour. The linear AEC alone provides limited and unstable echo attenuation, particularly under double-talk conditions, due to adaptation interruptions and an inability to track the time-varying acoustic environment. The addition

of a RES stage improves stability during single-talk and yields moderate gains in suppression performance; however, residual sensitivity to estimation errors and adaptation mismatches remains evident.

Extending the cascade with NS further increases objective ERLE by reducing additional non-echo energy components, but this comes at the cost of increased speech distortion and error propagation through the processing chain. In contrast, the proposed deep learning post-filter demonstrates consistently strong performance across both single-talk and double-talk conditions, maintaining high ERLE while exhibiting rapid recovery during acoustic transitions. Importantly, this performance is achieved despite inaccuracies in the underlying linear echo estimate, indicating improved robustness to front-end limitations.

Taken together, the results demonstrate a fundamental trade-off in traditional cascaded architectures between increasing suppression and preserving speech quality, whereas the learned post-filter approach is able to decouple this trade-off by directly modelling and suppressing residual interference in a more adaptive manner.

6.5 Computational Performance and Real-Time Considerations

In addition to objective and perceptual performance, practical deployment of AEC systems requires adherence to strict real-time constraints, particularly in embedded and automotive environments. These constraints are typically expressed in terms of algorithmic latency and RTF, which jointly determine whether a system can operate in an interactive full-duplex communication setting.

6.5.1 Latency Analysis

For real-time deployment, system latency must account for both signal analysis delays and streaming update behaviour. In block-based speech enhancement systems, latency arises

primarily from the analysis front-end and the temporal context required by the enhancement model.

The proposed hybrid model operates on a STFT front-end configured as follows:

- Sampling rate (f_s): 16 kHz
- Window length (N_w): 320 samples (20 ms)
- Hop length (N_h): 160 samples (10 ms)

STFT Algorithmic Latency The STFT front-end introduces a fixed buffering delay corresponding to the window length required to compute the first spectral frame. At a sampling rate of 16 kHz and a window length of $N_w = 320$ samples, this results in:

$$\text{Latency}_{\text{STFT}} = \frac{N_w}{f_s} = \frac{320}{16000} = 20\text{ms}. \tag{6.1}$$

This represents the fundamental signal processing latency introduced by time-frequency analysis and is independent of the subsequent neural processing stage.

Neural Streaming Latency The neural post-filter operates on a heavily overlapping context window with a stride of 2 frames. Due to the high overlap ratio, the effective additional algorithmic delay introduced by the neural processing pipeline corresponds to approximately one frame of dependency, yielding:

$$\text{Latency}_{\text{DL}} \approx 10\text{ms}. \tag{6.2}$$

Total Algorithmic Latency The resulting end-to-end algorithmic latency is therefore:

$$\text{Latency}_{\text{total}} = 20\text{ms} + 10\text{ms} = 30\text{ms}. \tag{6.3}$$

This value represents the worst-case signal path delay from audio acquisition to enhanced output in steady-state operation.

Streaming Inference Configuration The neural post-filter operates on a rolling context window of $C_{size} = 16$ (160 ms) frames with an overlap of $O = 14$ frames, resulting in a stride of:

$$S = C_{size} - O = 2 \text{ frames}. \tag{6.4}$$

Given a hop size of 10 ms per frame, the system produces updated outputs every:

$$\text{Update Interval} = S \cdot N_h = 2 \times 10\text{ms} = 20\text{ms}. \tag{6.5}$$

This update interval defines the temporal resolution at which enhanced estimates are produced in steady-state operation. It represents the output refresh rate of the system rather than an additional source of algorithmic latency.

Discussion This configuration achieves a favourable balance between temporal context and responsiveness. The model leverages 160 ms of contextual information while maintaining a 30 ms latency and 20 ms update interval, enabling robust tracking of late reflections and dynamic acoustic conditions while remaining suitable for real-time deployment.

6.5.2 RTF Evaluation

To evaluate computational feasibility, the RTF is used as a standard measure of processing efficiency:

$$\text{RTF} = \frac{\text{Total Execution Time}}{\text{Total Audio Duration}}. \tag{6.6}$$

A system is considered real-time if RTF < 1.0, meaning it processes audio faster than it is generated.

When evaluated on an Apple M1 CPU using an unoptimized Python implementation,

the proposed model achieved an RTF of approximately 0.68. This confirms that the system satisfies real-time requirements for online processing.

However, this value remains above the strict constraints typically imposed in production-grade embedded automotive systems, where RTF ≤ 0.5 is often required to ensure sufficient computational headroom for concurrent tasks such as diagnostics, networking, and control logic.

It is important to note that this result represents a conservative upper bound. The current implementation relies on a high-level Python/PyTorch inference pipeline on CPU, which does not reflect deployment-optimized conditions. Significant improvements are expected through standard optimization techniques, including model quantization, pruning, and conversion to optimized inference runtimes such as ONNX Runtime, as well as hardware acceleration.

Compared to traditional signal processing pipelines, which rely on lightweight fixed-function operations, the proposed hybrid model introduces higher computational complexity. This increase is a direct trade-off for improved perceptual performance and robustness. However, the measured RTF indicates that the system remains within a practical range for modern embedded architectures.

6.5.3 Discussion

The results demonstrate that the proposed hybrid model satisfies key real-time constraints while significantly improving enhancement performance. The 20 ms output update interval ensures responsive behaviour suitable for interactive full-duplex communication systems, while the causal design guarantees zero lookahead beyond past context.

Although the measured RTF of 0.68 exceeds the most aggressive embedded deployment targets, it remains within a realistic optimization range. With standard deployment optimizations, the model is expected to meet real-world automotive requirements.

Overall, these findings support the central hypothesis of this work: a unified neural architecture can outperform cascaded signal processing pipelines while still maintaining practical

real-time feasibility. The following section presents an ablation study to further isolate the contribution of individual architectural components.

6.6 Ablation Study

To better understand the contribution of individual design choices within the proposed hybrid deep learning framework, a comprehensive ablation study was conducted. The objective of this study is to isolate the impact of key architectural, representational, and training-related components and quantify their contribution to overall system performance.

Unlike the comparison with traditional cascaded approaches, which evaluates fundamentally different processing paradigms, the ablation study focuses exclusively on variations of the proposed model. Each experiment modifies a single aspect of the system while keeping all other components fixed, enabling a controlled analysis of design trade-offs.

This allows the final system design choices to be interpreted in terms of both performance contribution and computational/training trade-offs.

The evaluated variations span multiple aspects of the system design, including architectural configuration, mask representation, loss function formulation, input feature selection, temporal context modeling, and training strategy. In particular, the following components are investigated:

- Network architecture (e.g., number of F-T-LSTM layers, removal of recurrent components)
- Mask representation (complex ratio mask vs magnitude-based masking)
- Loss functions (different perceptual and signal-based objectives)
- Input features (use of estimated echo vs far-end reference signal)
- Temporal context and buffering strategy (frame window size and overlap)

- Stateful vs non-stateful processing
- Training hyperparameters (batch size, learning rate, optimizer)

To ensure clarity and interpretability, the ablation results are presented in a structured order that reflects the system design pipeline. The analysis begins with architectural variations, followed by mask representation, loss formulation, input configuration, and temporal modeling choices. Finally, training-related factors are examined.

This progression allows the impact of each design decision to be understood in the context of subsequent components, providing a coherent view of how the final system configuration was derived.

6.6.1 Architectural Variations

To investigate the trade-off between model complexity and performance, several architectural variations were evaluated throughout the development of the proposed system. The initial design was based on the architecture proposed by [15], which served as the starting point for this work. However, several modifications were introduced to better suit the residual echo suppression task and the requirements of real-time deployment. Specifically, the original near-end speech estimation objective was replaced with CRM estimation, the VAD branch was removed, and the dual decoder structure used for separate real and imaginary outputs was replaced with a single decoder due to the lower complexity of CRM prediction.

Starting from this modified reference architecture, additional simplifications were investigated to reduce computational complexity while maintaining performance. The primary architectural modification was the progressive reduction of decoder complexity through the removal of the final skip connection and its corresponding decoder stage. This modification also enabled a reduction in the maximum hidden dimensionality of the F-T-LSTM bottleneck from 256 to 128. In the reference architecture, the deepest encoder and decoder representations operated at a dimensionality of 256, requiring the F-T-LSTM layers to process feature

maps of the same size. Following the removal of the deepest decoder stage, the bottleneck instead interfaced directly with the next decoder level operating at a dimensionality of 128. Consequently, the maximum F-T-LSTM hidden size could be reduced accordingly, decreasing the computational cost of the recurrent layers while maintaining architectural consistency between the bottleneck and decoder. Furthermore, configurations with no F-T-LSTM, a single F-T-LSTM, and the full three-layer F-T-LSTM bottleneck were evaluated to determine the minimum temporal modeling capacity required for effective residual echo suppression.

Reference Model

The reference model retains the full U-Net style encoder-decoder structure and represents the architecture most closely aligned with [15] after the task-specific modifications described above. This configuration serves as the starting point from which subsequent architectural simplifications were investigated. From an AECMOS perspective, the reference model demonstrates strong echo suppression performance across all scenarios. In double-talk conditions, the model achieves an Echo MOS of 4.69, indicating excellent residual echo attenuation. However, this performance comes at the expense of speech preservation, with the corresponding Degradation MOS falling to 1.96. This suggests that the model prioritizes aggressive echo suppression, occasionally introducing audible distortion to the near-end speech signal.

In far-end single-talk conditions, where echo suppression is the primary objective, the model achieves an Echo MOS of 4.14 while maintaining a near-perfect Degradation MOS of 4.99. Similarly, in near-end single-talk conditions, where no echo is present, the model preserves the maximum Echo MOS score of 5.00 while achieving a Degradation MOS of 3.34. A similar trend is observed in the DNSMOS results. The model achieves Overall scores of 3.85, 3.78, and 4.05 for double-talk, far-end single-talk, and near-end single-talk conditions, respectively, indicating strong perceptual quality while maintaining effective residual echo suppression. Only the overall scores were evaluated for brevity. These results are highly

comparable to those of the final architecture. The primary difference is that the reference model exhibits a stronger emphasis on echo suppression at the expense of speech preservation, whereas the final model achieves a more balanced trade-off between residual echo attenuation and speech quality.

Despite these results, the reference model exhibits significant computational complexity, with an average RTF of 1.42, making real-time deployment challenging. To address this limitation, the final skip connection and its corresponding decoder stage were removed, producing the architecture adopted for the proposed system.

The resulting model achieves nearly identical objective performance, with only minor differences in both AECMOS and DNSMOS metrics. In particular, the simplified architecture exhibits a slight shift toward improved speech preservation, reflected by higher Degradation MOS scores, while maintaining comparable echo suppression performance. The negligible performance difference suggests that the removed decoder stage contributes limited useful information to the CRM estimation task. Unlike the original architecture in [15], which reconstructs the complete near-end speech signal, the proposed system estimates a suppression mask applied to the output of the linear AEC, resulting in a substantially simpler reconstruction problem. Furthermore, the F-T-LSTM bottleneck already provides strong contextual modeling and feature propagation, reducing the dependence on the deepest encoder-decoder skip pathway.

Most importantly, the architectural reduction decreases the average RTF from 1.42 to 0.68 while maintaining nearly identical objective performance. Although training convergence becomes slightly slower, requiring approximately 100–150 epochs compared to 70–80 epochs for the reference model, the substantial reduction in runtime complexity together with the preservation of perceptual quality strongly justifies the adoption of the simplified decoder structure.

While the decoder reduction produced substantial computational savings with minimal performance impact, it remained unclear whether the F-T-LSTM bottleneck was still neces-

sary. Therefore, additional architectural variants were evaluated by removing the F-T-LSTM layers entirely.

Model Without F-T-LSTM

Removing the F-T-LSTM component results in a purely convolutional architecture, where temporal dependencies are captured only through the finite receptive field of the convolutional layers.

From an AECMOS perspective, this configuration demonstrates limited echo suppression capability. In double-talk conditions, the model achieves moderate performance (Echo MOS of 4.14 and Degradation MOS of 2.19), indicating that local time–frequency features are sufficient to partially suppress residual echo. However, in far-end single-talk conditions, performance degrades significantly, with Echo MOS dropping to 2.91, highlighting the model's inability to effectively model long-term echo structures.

In near-end single-talk conditions, the model maintains an Echo MOS of 5.00, as expected, but exhibits reduced Degradation MOS (2.76), suggesting that the absence of temporal modeling leads to unnecessary distortion of clean speech.

A similar trend is observed in DNSMOS. The model achieves moderate performance in double-talk conditions (3.41), but underperforms in far-end single-talk (3.20), further confirming its limited ability to suppress sustained echo components. Near-end performance remains relatively strong (4.03), as minimal processing is required in this scenario.

These results indicate that while convolutional layers can capture local spectral patterns, they are insufficient for modeling the temporal dynamics required for effective echo cancellation, particularly in scenarios dominated by far-end signals.

Single F-T-LSTM Layer

Introducing a single F-T-LSTM layer significantly improves performance across all scenarios, demonstrating the importance of explicit temporal modeling.

In double-talk conditions, the model achieves strong echo suppression (Echo MOS of 4.66), albeit with a reduction in Degradation MOS (1.97), indicating a more aggressive suppression strategy. In far-end single-talk conditions, performance improves substantially (Echo MOS of 4.08 and Degradation MOS of 5.00), reflecting the model's enhanced ability to track and cancel sustained echo components. Near-end performance also improves compared to the convolution-only model, with a Degradation MOS of 3.20.

From a perceptual standpoint, the DNSMOS results further confirm this improvement. The model achieves strong overall scores across all scenarios, with 3.75 in double-talk, 3.35 in far-end single-talk, and 3.94 in near-end single-talk. These results indicate that even a single F-T-LSTM layer is sufficient to significantly improve perceptual quality compared to the convolution-only configuration.

Overall, the single F-T-LSTM configuration achieves AECMOS performance comparable to the final three-layer design, indicating that even a limited amount of temporal modeling is sufficient to effectively suppress residual echo. However, consistent reductions in DNSMOS scores are observed across the evaluated scenarios (compared to final model), suggesting that the additional temporal context provided by the deeper F-T-LSTM bottleneck contributes to improved perceptual speech quality. This is particularly reasonable given that both residual echo and background noise often exhibit long-term temporal structure that can benefit from extended temporal modeling.

A more significant difference is observed during training. While the final three-layer architecture typically converges within approximately 100–150 epochs, the single F-T-LSTM configuration requires nearly 500 epochs to achieve comparable performance. Despite providing a good RTF improvement of approximately 0.46, the substantially slower convergence indicates that the reduced temporal modeling capacity makes optimization more difficult and limits the robustness of the learned representations. As a result, the three-layer F-T-LSTM design was selected as it provides improved perceptual quality and significantly better training efficiency while incurring only a minor increase in runtime complexity.

Three F-T-LSTM Layers (Proposed Configuration)

The three-layer F-T-LSTM architecture was selected as the final configuration due to its favorable balance between performance, training efficiency, and runtime complexity. While the single F-T-LSTM model achieves comparable AECMOS performance, the deeper architecture consistently produces higher DNSMOS scores, indicating improved perceptual speech quality resulting from the additional temporal modeling capacity.

A more significant distinction is observed during training. The three-layer configuration typically converges within approximately 100–150 epochs, whereas the single-layer model requires nearly 500 epochs to achieve comparable performance. This suggests that the additional F-T-LSTM layers provide more robust temporal representations that facilitate optimization and improve learning efficiency.

Although the single-layer model offers a decent reduction in RTF, the runtime improvement is not worth the substantial increase in training time and the reduction in perceptual quality. This is why the three-layer F-T-LSTM configuration was selected for the final model as it provides the most favorable trade-off between performance, computational cost, and training stability.

6.6.2 Mask Representation

To evaluate the impact of output representation, the proposed CRM formulation was compared against a magnitude-only IRM masking approach. In the magnitude mask configuration, the model estimates a real-valued mask applied only to the magnitude of the STFT, while the phase of the input signal is left unchanged.

This is particularly relevant in echo cancellation scenarios, where phase alignment between microphone and echo reference signals plays a critical role due to propagation delays and room reflections.

From an AECMOS perspective, the magnitude mask configuration results in consistently worse performance across all scenarios. In double-talk conditions, the model achieves an

Echo MOS of 4.53 and Degradation MOS of 2.06, both lower than the corresponding CRM-based model. In far-end single-talk conditions, Echo MOS decreases to 3.66 (compared to 3.96 with CRM), while Degradation MOS remains high at 5.00. In near-end single-talk conditions, although Echo MOS remains at 5.00, Degradation MOS drops significantly to 2.29, indicating increased distortion of the near-end speech signal.

A similar trend is observed in the DNSMOS results. Considering overall scores, the magnitude mask model achieves 3.60 in double-talk, 3.15 in far-end single-talk, and 3.92 in near-end single-talk. In all cases, these values are lower than those obtained using the CRM-based model.

These results highlight a key limitation of magnitude-only masking approaches. By operating exclusively on the magnitude spectrum, the model is unable to modify the phase component of the signal, which can impact the accuracy of the reconstruction in the time-domain. In AEC scenarios, both echo and noise components exhibit phase-dependent characteristics, particularly due to propagation delays, reflections, and non-linear distortions in the acoustic path.

As a result, magnitude-only masking can lead to suboptimal suppression and introduce artifacts during signal reconstruction, especially in complex scenarios such as double-talk. In contrast, the CRM formulation enables joint modeling of both magnitude and phase, allowing the network to more accurately estimate the underlying clean signal and achieve improved perceptual quality.

Overall, the CRM-based representation provides a clear advantage over magnitude-only masking, and is therefore adopted in the final model configuration.

6.6.3 Loss Function Design

The choice of loss function plays an important role in balancing echo suppression and speech preservation. In the proposed framework, multiple loss formulations were evaluated to determine the most effective objective for training the CRM-based model.

Training Stability Considerations

It was observed that training with a complex-valued objective is more sensitive than magnitude-only formulations. In particular, directly optimizing a full CRM-based loss from random initialization often resulted in unstable convergence during the early stages of training. Two different strategies were tested and used to improve optimization stability. First, a staged training approach was adopted in which early training emphasizes magnitude-based loss components before gradually increasing the contribution of the complex-domain objectives. This allows the network to first learn a coarse spectral representation before being tasked with the more challenging phase-aware optimization. Second, a learning-rate warm-up schedule was utilized, whereby the learning rate is gradually increased from a small initial value to the target learning rate before transitioning to the cosine annealing schedule. This reduces the likelihood of large parameter updates during the initial epochs and further improves training stability. Overall, these techniques were found to significantly improve convergence when training with the proposed CRM-based objective.

This behaviour reflects the increased difficulty of jointly estimating magnitude and phase, as errors in phase estimation can significantly impact time-domain reconstruction.

Hybrid Magnitude and CRM Loss

An initial approach combined magnitude mask loss with CRM supervision using the clean near-end signal as a reference. While this hybrid formulation provided a reasonable baseline, it resulted in suboptimal performance across all evaluation metrics.

This suggests that directly combining magnitude and complex-domain objectives without careful balancing can lead to competing gradients, which limits the model's ability to converge to an optimal representation.

From an AECMOS perspective, the model achieved Echo MOS of 4.52 (double-talk), 3.74 (far-end), and 5.00 (near-end), with corresponding Degradation MOS values of 2.07, 5.00, and 2.93. These results indicate moderate echo suppression but noticeable distortion

in near-end conditions.

Similarly, DNSMOS overall scores were limited, with 3.646 (double-talk), 3.449 (far-end), and 3.908 (near-end), confirming that the model struggled to achieve strong perceptual quality using this formulation alone.

Incorporating Reconstruction Loss

To improve alignment with perceptual quality, a reconstruction-based loss component was introduced. This type of loss rebuilds the signal after the mask is applied and compares it to the ground-truth near-end signal in the time domain. This modification resulted in consistent performance improvements across both AECMOS and DNSMOS.

Echo suppression improved in far-end conditions (Echo MOS increased from 3.74 to 3.99), while near-end degradation was reduced (Degradation MOS improved from 2.93 to 3.17). DNSMOS scores also showed modest gains, particularly in double-talk and near-end scenarios.

This indicates that incorporating time-domain reconstruction constraints helps align the optimization objective with perceptual quality metrics, leading to more stable and consistent enhancement behaviour.

These improvements suggest that explicitly constraining the reconstructed signal helps the model produce more perceptually consistent outputs, rather than relying solely on mask-level supervision.

Final Loss Formulation

The best performance was obtain using the composite loss introduced in 4.4, defined as:

$$\begin{aligned}\mathcal{L}_{\text{total}} =& \alpha\, \mathcal{L}_{\text{cplx}} + (1-\alpha)\, \mathcal{L}_{\text{mag}} \\ &+ \lambda_{\text{SI-SDR}}\, \mathcal{L}_{\text{SI-SDR}} + \lambda_{\text{bias}}\, \mathcal{L}_{\text{bias}}.\end{aligned} \tag{6.7}$$

This multi-objective loss framework balances spectral-domain supervision, time-domain per-

ceptual alignment, and task-specific structural biasing to optimize the model across contrasting acoustic environments.

Hyperparameter Sensitivity Analysis

To isolate the explicit influence of the loss components and optimization bounds, a sensitivity analysis was conducted by varying key hyperparameters. Several parameters evaluated below are embedded directly within the individual sub-loss formulations of the composite framework.

Asymmetry Factor ($\alpha_{\mathbf{asym}}$) The asymmetry factor dictates how heavily the spectral losses ($\mathcal{L}_{\text{cplx}}$ and $\mathcal{L}_{\text{mag}}$) penalize under-suppression of echo versus over-suppression of speech. Minimizing this factor from the baseline consistently degraded both far-end and double-talk metrics. Without an asymmetric bias, the network treats residual echo leaks and minor near-end target speech energy variations with equal weight. Forcing the loss to be symmetric causes the model to under-attenuate harsh echo leaks during double-talk bursts, confirming that an intentional asymmetry is helpful to enforce aggressive echo suppression.

SI-SDR Weight ($\lambda_{\mathbf{SI\text{-}SDR}}$) Scaling up the time-domain SI-SDR loss weight yielded cleaner structural echo reduction but triggered a sharp drop in the near-end Degradation AECMOS. Because SI-SDR enforces rigid, sample-level phase alignment in the time domain, over-emphasizing it forces the network to prioritize exact sample matching over broader perceptual preservation. This strict mathematical constraint introduces micro-alignment errors and musical noise artifacts in the STFT domain, highlighting a clear boundary where aggressive time-domain optimization begins to actively damage speech naturalness.

Spectral Weighting (α) The α parameter controls the balance between complex-domain phase tracking ($\mathcal{L}_{\text{cplx}}$) and pure magnitude reconstruction ($\mathcal{L}_{\text{mag}}$). Tilting the weight heavily toward the complex spectrum initially sharpened echo path modeling, but ultimately

introduced noticeable distortion in near-end single-talk scenarios. This behavior suggests that over-constraining the model to optimize complex phase relationships can cause the magnitude reconstruction to fluctuate erratically. The resulting phase-modulation artifacts manifest as a subtle "gargling" effect that lowers perceptual quality when the local speaker talks in isolation.

Mask Bias Weight (λ_{bias}) Increasing the mask bias weight forces the estimated CRM coefficients closer to zero in bins containing high estimated echo energy. While this structural bias successfully wiped out residual echo during double-talk and far-end single-talk, it heavily penalized near-end speech quality. The aggressive masking mechanism struggles with spectral resolution limits; as a result, it treats weak, unvoiced near-end consonants and transient speech onsets as echo residuals, clipping them out entirely and leaving behind a fractured, unnatural speech envelope.

Exponent Parameter (p) The exponent parameter controls the power-law compression applied to the spectrogram magnitudes to mimic human auditory loudness perception. Deviating from the optimal baseline value degraded performance across all AECMOS metrics. Setting p too high (closer to a linear scale) caused loud echo peaks to completely dominate the gradient updates, leaving quiet speech components unoptimized. Conversely, lowering p too far overly compressed the dynamic range, causing the network to over-emphasize low-level background noise floors at the expense of tracking primary voice characteristics. This sensitivity confirms that the selected exponent provides the ideal gradient balance across diverse signal amplitudes.

Summary

Overall, the ablation study shows that effective loss design requires balancing spectral, time-domain, and biasing objectives. Purely spectral or purely time-domain losses are insufficient to achieve consistent perceptual quality.

The final formulation provides a balanced trade-off between echo suppression, speech preservation, and perceptual quality, and is therefore used in the proposed model.

6.6.4 Input Feature Configuration

To evaluate the impact of input representation, an ablation study was conducted comparing the use of the far-end reference signal versus the estimated echo signal as an input to the model. This design is motivated by prior work [15], which shows that the echo estimate provides more structured and task-relevant information than the raw far-end signal.

Accordingly, two input configurations were evaluated:

- **Far-end configuration:** microphone signal, far-end reference, and linear AEC output
- **Echo estimate configuration:** microphone signal, estimated echo signal, and linear AEC output

Performance Comparison

Across double-talk and far-end single-talk scenarios, both configurations achieve similar echo suppression performance, suggesting that the model can learn to suppress residual echo from either representation.

A more pronounced difference appears in near-end single-talk conditions, where the far-end input leads to increased distortion (lower Degradation MOS), whereas the estimated echo improves near-end speech preservation.

Discussion

This can be explained by the nature of the inputs. The far-end signal is only indirectly related to the observed echo due to unknown acoustic effects such as delay, reflections, and non-linearities, requiring the model to implicitly learn this transformation.

In contrast, the estimated echo already incorporates this transformation through the linear AEC stage, providing a structured approximation of the echo component. This reduces the learning burden and allows the model to focus on residual echo and non-linear artifacts.

This aligns with [15], which reports improved performance when using echo-informed inputs.

Summary

Overall, both input configurations achieve comparable echo suppression, but the estimated echo provides more consistent near-end speech preservation and is therefore used in the final model.

6.6.5 Temporal Context and Sliding Window Configuration

To balance performance with real-time constraints, several sliding window configurations were evaluated using overlapping STFT frames. The window size determines the available temporal context, while the overlap controls the update rate of the streaming system.

The evaluated configurations include 4/2, 16/8, and 16/14, representing different trade-offs between temporal context, computational cost, and responsiveness.

Small Context (4/2)

Using a small window of 4 frames with an overlap of 2 frames results in minimal temporal context and reduced computational cost, but significantly degraded performance. From an AECMOS perspective, while Echo MOS remains relatively high in double-talk (4.69) and far-end conditions (3.96), Degradation MOS drops sharply to 1.40 (double-talk) and 1.47 (near-end), indicating severe distortion of the near-end speech signal.

This behaviour is also evident in subjective listening, where outputs are noticeably unstable and over-suppressed.

The limitation is insufficient temporal context. Although the F-T-LSTM is stateful, the convolutional encoder relies on a limited temporal receptive field and cannot adequately model long-duration echo components, particularly under reverberant conditions.

Large Context Configurations

Increasing the window size to 16 frames significantly improves performance by providing sufficient temporal context for both convolutional and recurrent components.

Two overlap configurations were evaluated:

16/14 (Proposed Configuration) This configuration uses a stride of 2 frames. Since each frame corresponds to 10 ms, the system produces updated outputs every:

$$\text{Update Interval} = 2 \times 10\,\text{ms} = 20\,\text{ms}. \tag{6.8}$$

At system initialization, a short warm-up period is required to populate the 16-frame context window.

The 16-frame window corresponds to a temporal context of 160 ms, which enables robust modelling of long-term echo dynamics without affecting steady-state responsiveness.

This configuration provides strong performance across all evaluation metrics and is used in the final model.

16/8 Reducing the overlap to 8 frames increases the stride to 8 frames, resulting in a slower update rate:

$$\text{Update Interval} = 8 \times 10\,\text{ms} = 80\,\text{ms}. \tag{6.9}$$

While this configuration reduces computational cost and improves runtime efficiency (lower RTF ≈ 0.20 on M1 hardware), it introduces a significantly slower output refresh rate, making it less suitable for real-time interaction.

Summary

These results highlight the importance of temporal context in echo cancellation. Small contexts lead to instability and distortion, while larger contexts improve modelling of long-term echo structure.

However, increasing the stride reduces the update rate of the system, introducing a trade-off between responsiveness and computational efficiency. The selected 16/14 configuration provides an effective balance, achieving strong performance while maintaining a 20 ms steady-state update interval suitable for real-time deployment.

6.6.6 Stateful vs Non-Stateful Processing

To evaluate the importance of temporal memory beyond the input window, a comparison was conducted between stateful and non-stateful versions of the proposed model. In the stateful configuration, hidden states of the F-T-LSTM layers are preserved across consecutive input segments, whereas in the non-stateful configuration, hidden states are reset at each processing step.

Performance Comparison

The non-stateful model exhibits consistently worse performance across all evaluation metrics. From an AECMOS perspective, Echo MOS decreases across all scenarios, with values of 4.22 (double-talk), 3.56 (far-end), and 5.00 (near-end). Near-end Echo MOS remains unchanged (5.00), as expected, while performance decreases in double-talk and far-end conditions.

Degradation MOS also shows a decline, particularly in near-end conditions, where it drops to 3.14, indicating increased distortion of clean speech.

A similar trend is observed in DNSMOS results. Considering overall scores, the non-stateful model achieves 3.567 (double-talk), 3.322 (far-end), and 3.938 (near-end), all of which are lower than the corresponding scores of the stateful configuration. This indicates reduced perceptual quality and less effective noise suppression.

Summary

Overall, the stateful configuration provides clear advantages in both objective and perceptual performance. By preserving temporal information across frames, the model is better equipped to handle long-duration echo and noise components, leading to improved echo suppression and reduced signal distortion. As a result, the stateful formulation is adopted in the final model.

6.6.7 Training Hyperparameters

The final model uses a batch size of 8 and a learning rate of 2.5×10^{-4} with AdamW and weight decay 1×10^{-4}.

Batch Size and Learning Rate

Batch sizes from 4 to 32 were evaluated. Small batches produced noisy gradients and unstable training, while large batches can be detrimental to robust convergence and generalization. A batch size of 8 provided the best trade-off.

The learning rate was selected in conjunction with the batch size, following standard scaling considerations, where larger batch sizes typically require increased learning rates to maintain effective gradient magnitudes. The chosen value of 2.5×10^{-4} was found to provide stable and efficient convergence for the selected batch size.

Optimizer Selection

The AdamW optimizer was selected due to its ability to decouple weight decay from gradient-based updates, improving generalization compared to standard Adam. Alternative optimizers were considered but did not provide consistent improvements.

Summary

The selected hyperparameters ensure stable convergence and strong generalization, but have less impact on final performance than architectural and loss design choices.

The results presented in this chapter provide a comprehensive evaluation of the proposed system across objective metrics, temporal behaviour, computational constraints, and architectural design choices. The following section discusses these findings in a broader context, highlighting key trends, trade-offs, and their implications for real-world acoustic echo cancellation systems.

6.7 Discussion

The results in this chapter provide a comprehensive evaluation of the proposed hybrid deep learning AEC system across perceptual quality, temporal behaviour, signal-level performance, and computational efficiency. This section synthesizes these findings and discusses their implications for system design and real-world deployment.

6.7.1 Performance Trade-offs in Echo Cancellation

A consistent trend across all experiments is the trade-off between echo suppression and speech preservation. The reference real-world cascaded system used for comparison achieved strong echo reduction performance, particularly through the combined use of AEC, AES, and NS stages. However, this often introduced additional speech distortion, reflected in lower Degradation MOS scores and reduced perceptual speech quality metrics.

In contrast, the proposed hybrid model achieved a more balanced outcome. While overall echo suppression performance remained competitive with the traditional cascaded approach, the proposed model consistently demonstrated improved speech preservation and perceptual quality across most evaluated scenarios. These improvements were particularly noticeable under double-talk conditions, where simultaneous near-end and far-end speech create more

challenging acoustic environments.

This behaviour is likely related to the ability of the deep learning post-filter to better model residual non-linear echo components and complex distortions that remain after linear echo cancellation. Traditional adaptive filtering approaches are highly effective for linear echo suppression but are generally less capable of handling non-linearities introduced by loudspeakers, microphones, and acoustic propagation effects. As a result, the proposed hybrid model was able to maintain similar suppression performance while reducing perceptual speech degradation.

The DNSMOS evaluation further supports these observations. In addition to improved speech preservation, the proposed approach generally achieved stronger perceptual noise suppression performance across most tested scenarios. This indicates that the jointly optimized enhancement model was able to simultaneously suppress residual echo and background noise more effectively than the independently optimized cascaded processing pipeline.

These improvements stem from joint optimization within a single model. Unlike cascaded pipelines, where each stage is independently optimized for a local objective, the proposed approach learns all enhancement objectives simultaneously through a composite loss. This avoids conflicting processing behaviours, where aggressive suppression in one stage can degrade the input for subsequent stages.

The shared representation further enables the model to jointly encode relationships between echo, speech, and noise, leading to more consistent behaviour across acoustic conditions. Overall, this reduces inter-stage error propagation and results in a more globally optimal solution.

6.7.2 Temporal Modelling and Dynamic Behaviour

Temporal modelling plays a critical role in system performance. Configurations with limited temporal context or without recurrent structure show degraded performance, particularly in far-end conditions where echo exhibits long-duration dependencies.

The inclusion of F-T-LSTM layers improves the ability to model these dependencies, resulting in stronger echo suppression and more stable behaviour. This is further enhanced by the stateful formulation, which extends temporal memory beyond the fixed input window.

The importance of temporal modelling becomes especially apparent during double-talk periods, where the system must simultaneously preserve near-end speech while suppressing residual echo. The recurrent bottleneck allows the model to better track temporal acoustic patterns and maintain more stable enhancement behaviour during rapid transitions between far-end only and double-talk conditions.

The ERLE analysis further reinforces these observations. The traditional system achieves lower overall echo suppression and exhibits noticeable performance degradation during double-talk periods, resulting in reduced ERLE values and slower recovery following double-talk events. This behaviour is largely attributed to the operation of the DTD, which temporarily suspends adaptation to prevent filter divergence. Consequently, the adaptive filter requires additional time to re-converge once double-talk has ended. In contrast, the proposed deep learning approach maintains more consistent suppression throughout these transitions, exhibiting faster recovery and more stable ERLE behaviour across changing acoustic conditions. These results suggest improved robustness in dynamic environments where periods of single-talk and double-talk frequently alternate.

6.7.3 Advantages of the Hybrid Learning Approach

The hybrid architecture combines a linear AEC front-end with a deep learning post-filter. The linear stage removes dominant linear echo components, while the neural network focuses on residual echo, non-linearities, and noise.

This separation simplifies the learning task and improves efficiency. Rather than learning full echo cancellation from raw signals alone, the neural network operates on a partially enhanced signal where dominant linear echo components have already been removed. This allows the deep learning stage to focus computational capacity on more difficult residual

artifacts and perceptual enhancement.

The ablation study further confirms that using the estimated echo signal provides more structured input than the raw far-end reference, improving near-end speech preservation and enhancement stability. The estimated echo signal provides direct information about residual echo structure, enabling the network to better distinguish between near-end speech and unwanted echo components.

The use of a CRM output further enhances performance by enabling joint modelling of magnitude and phase, which is critical for accurate signal reconstruction in complex acoustic environments. Compared to magnitude-only masking approaches, the CRM formulation improves reconstruction quality and contributes to the improved perceptual metrics observed throughout the evaluation.

From a deployment perspective, the masking-based formulation offers additional advantages over direct signal estimation approaches such as [15]. By applying a learned mask to the microphone signal rather than generating a full waveform, the model reduces output complexity and improves robustness to distribution shifts. This leads to more stable generalization in unseen acoustic conditions while maintaining computational efficiency.

6.7.4 Real-Time Considerations and Practical Deployment

The proposed model achieves a latency of 30 ms and an RTF of approximately 0.68 in a Python implementation. While this does not meet all strict benchmark constraints, it is close to real-time operation and can be further improved through optimized implementations in lower-level languages such as C++. Larger contexts were found to improve enhancement performance and temporal consistency, while smaller contexts produced noticeable degradation in suppression quality and perceptual stability.

Importantly, the proposed architecture demonstrates that strong enhancement performance can be achieved while remaining close to real-time constraints. Many existing deep learning approaches prioritize enhancement quality without considering deployment limita-

tions, while traditional production systems often sacrifice perceptual performance for computational simplicity. The proposed hybrid framework attempts to balance both requirements by combining efficient adaptive filtering with a compact temporally-aware deep learning architecture.

It is also important to note that the benchmark constraints referenced in this work (ICASSP challenge settings) are often stricter than those required in real-world applications. As such, the proposed system is already viable in scenarios with moderate RTF tolerance.

6.7.5 Overall System Insights

Overall, the results show that effective AEC design requires balancing temporal modelling, representation choice, loss formulation, perceptual quality, and computational constraints. The proposed hybrid system integrates signal processing and deep learning in a way that leverages the strengths of both approaches.

The experimental results demonstrate that the proposed system is capable of achieving echo suppression performance comparable to a practical real-world cascaded AEC system while improving speech preservation and perceptual quality. The improvements observed during double-talk conditions and noise suppression evaluations further highlight the advantages of combining adaptive filtering with learned residual enhancement.

Performance improvements arise not from a single component, but from the interaction of multiple design choices, including CRM-based output, stateful temporal modelling, structured hybrid processing, and joint loss optimization. This highlights the importance of system-level design rather than isolated component optimization.

In summary, the proposed model demonstrates competitive echo suppression, improved speech preservation, stronger perceptual noise suppression, robustness to dynamic conditions, and promising real-time characteristics, making it a strong candidate for practical next-generation AEC systems.

Chapter 7

Conclusion and Future Work

7.1 Conclusion

This thesis focused on developing a hybrid deep learning and signal processing framework for real-time AEC. Modern AEC systems must simultaneously suppress echo, preserve near-end speech, remove background noise, and satisfy strict latency constraints under dynamic acoustic conditions. To address the gaps in existing literature, this work establishes a controlled evaluation framework that isolates the true performance of the neural post-filter from the underlying linear adaptive filter. This approach ensures that performance gains are directly attributable to the neural enhancement stage rather than variations in the front-end configuration or idealized simulation setups.

The proposed system combines a production-oriented linear PBFDAF front-end with a deep learning post-filter, forming a hybrid AEC architecture designed for deployment-oriented evaluation. The linear stage adopts a proprietary configuration inspired by the architecture in [46], explicitly chosen to reflect real-world implementation constraints rather than idealized adaptive filtering setups commonly used in prior work. The neural enhancement stage is based on the convolutional recurrent architecture from [15], but is redesigned to align with the proposed CRM-based formulation and real-time processing requirements. Instead of directly estimating the clean waveform, the network operates on the microphone

signal, linear residual error, and estimated echo signal to predict a CRM [21], thereby framing enhancement as structured post-filtering rather than full signal reconstruction. This design shifts the learning burden toward residual suppression and phase-aware refinement of the already partially cancelled signal. To better match this formulation and reduce architectural redundancy, the original dual-stream real/imaginary decoding pathway is replaced with a unified decoder that jointly estimates complex mask components. In addition, the auxiliary VAD branch is removed, and the decoder is further simplified through elimination of the final skip connection and corresponding decoding stage. These modifications reduce architectural complexity while improving consistency with the masking-based objective and supporting low-latency deployment constraints.

Experimental evaluation demonstrated that the proposed hybrid approach achieved echo suppression performance comparable to the reference real-world AEC system used throughout this work, while providing distinct improvements in speech preservation and perceptual quality. The reference system, which uses a traditional cascaded pipeline consisting of separate AEC, AES, and NS stages, achieved strong echo suppression across many scenarios. However, because both systems operated under an identical linear adaptive front-end, the evaluation confirmed that the proposed hybrid model consistently produced lower speech distortion and improved perceptual quality metrics, particularly under challenging double-talk conditions.

Another observation from the results is that while overall echo cancellation performance between the two systems is comparable, the primary advantages of the proposed approach are not found in raw linear echo reduction but in improved robustness during double-talk and reduced speech distortion. This is directly attributed to the ability of the deep learning post-filter to better model residual non-linear echo components and complex acoustic distortions that are not fully captured by linear adaptive filtering approaches [5]. In contrast, the cascaded system relies on independently optimized stages, which can lead to over-suppression or distortion when aggressive noise or echo suppression is applied sequentially.

The proposed approach also demonstrated strong performance for noise suppression across most evaluated scenarios. Improvements in DNSMOS scores indicate that the hybrid architecture was able to more effectively suppress background noise while maintaining perceptual speech quality. This suggests that the learned post-filter is not only compensating for residual echo but also acting as a more effective perceptual noise suppressor compared to traditional rule-based or independently trained noise suppression modules.

The results further demonstrated the importance of temporal modelling for AEC. The inclusion of F-T-LSTM layers and stateful temporal processing significantly improved the ability of the system to model long-term acoustic dependencies and maintain stable suppression performance under far-end and double-talk conditions. In addition, the use of the estimated echo signal as an additional input improved near-end speech preservation by providing a more structured representation of residual echo components compared to the raw far-end reference signal.

The proposed CRM-based masking formulation also provided advantages over direct signal estimation approaches. By learning a complex spectral mask applied to the microphone signal rather than directly generating a waveform as seen in challenge-tailored configurations [15], the system reduced output complexity while improving generalization and reconstruction stability under out-of-distribution acoustic conditions commonly encountered in real-world deployments. This improved generalization is especially important for practical AEC systems, where variations in room acoustics, speaker characteristics, and noise profiles can significantly degrade the performance of directly regressive approaches.

From a practical deployment perspective, the proposed system achieved promising real-time characteristics, maintaining a strict algorithmic latency of approximately 30 ms to satisfy automotive environment criteria, alongside an RTF near 0.68 within a Python-based implementation. Although the system does not fully satisfy all embedded production constraints, the results indicate that the architecture is already close to real-time execution limits and can likely be further optimized through lower-level language implementations,

model compression, and hardware-aware compilation. Furthermore, the benchmark constraints considered in this work are more restrictive than those encountered in many real-world communication systems, making the proposed approach practical for a wide range of applications [92].

Several limitations remain within the current work. The model was not specifically optimized for embedded deployment, and no pruning, quantization, or hardware-aware optimization techniques were applied. In addition, the implementation was developed primarily in Python, which limits achievable runtime efficiency compared to optimized C++ or embedded implementations. The training and evaluation conditions also remain constrained to the selected datasets and acoustic scenarios, and additional testing under broader real-world conditions would further validate generalization performance [93].

Overall, this thesis demonstrates that hybrid AEC systems integrating adaptive filtering with deep learning post-filtering can effectively combine the strengths of both approaches when framed within a controlled, deployment-aligned environment. The proposed architecture achieved competitive echo suppression performance relative to a practical, production-grade AEC system while improving speech preservation, perceptual quality, and noise suppression performance. These results highlight the importance of unified baseline benchmarking and targeted perceptual optimization in next-generation AEC systems operating under real-time constraints.

7.2 Future Work

While the proposed hybrid AEC system demonstrates strong performance across perceptual quality, speech preservation, and real-time operation, several opportunities remain for further improvement and extension.

One important direction for future work is deployment-oriented optimization. Although the proposed system achieved near real-time performance within a Python implementa-

tion, additional optimization techniques such as model pruning, quantization, knowledge distillation [93], and hardware-aware neural architecture design could significantly reduce computational complexity and latency. Furthermore, implementing the system in optimized lower-level languages such as C++ or deploying on embedded inference frameworks [92] could further improve RTF performance and make the architecture more suitable for commercial embedded systems.

Another promising direction involves improving robustness and generalization under more diverse acoustic conditions. Additional training and evaluation using larger real-world datasets containing varying reverberation levels, speaker characteristics, background noise conditions, and non-linear acoustic distortions would provide a more comprehensive assessment of practical deployment performance. Domain adaptation and online adaptation strategies could also be explored to improve performance under unseen acoustic environments.

An additional area of interest is the use of deep learning for improved estimation of echo and noise characteristics within practical communication systems. Traditional cascaded AEC, AES, and NS pipelines rely heavily on accurate residual echo and noise estimation, which can be difficult to achieve using conventional signal processing approaches alone, particularly under non-linear, double-talk, or rapidly changing acoustic conditions.

Future hybrid architectures could therefore be restructured to explicitly estimate residual echo and background noise components using learned representations [94]. These estimates could then be integrated into conventional AES and NS frameworks to improve suppression control, adaptation stability, and perceptual quality. Such an approach could combine the interpretability and robustness of traditional signal processing systems with the modelling capabilities of deep learning, potentially improving the performance of practical real-world enhancement pipelines.

Finally, future work could investigate extending the proposed framework to support multi-audio-channel and multi-microphone input configurations commonly used in practical communication systems [8]. Many real-world hands-free and in-vehicle AEC systems

utilize multiple microphones to improve speech capture, spatial awareness, and robustness under challenging acoustic conditions.

Incorporating multi-channel microphone inputs into the proposed hybrid framework could provide additional spatial and temporal information that may improve residual echo suppression, noise reduction, and near-end speech preservation. Multi-microphone processing could also improve robustness under double-talk conditions and highly reverberant environments where single-channel approaches become more limited.

Future architectures may therefore explore multi-channel temporal modelling, microphone fusion strategies, or integration with beamforming front-end processing [95] to better exploit spatial acoustic information. Such extensions would improve the practicality of the proposed system for deployment in modern conferencing, automotive, and smart device applications where multi-microphone configurations are standard.

Overall, the proposed work establishes a strong foundation for hybrid signal processing and deep learning approaches in AEC. Continued research in deployment optimization, learned signal estimation, perceptual modelling, and multi-channel audio processing has the potential to further advance the practicality, robustness, and performance of real-time speech enhancement systems in real-world communication applications.

References

[1] B. Farhang-Boroujeny, *Adaptive filters: theory and applications*. John Wiley & Sons, 2013.

[2] J. J. Shynk, "Frequency-domain and multirate adaptive filtering," *IEEE Signal processing magazine*, vol. 9, no. 1, pp. 14–37, 1992.

[3] P. C. Loizou, *Speech enhancement: theory and practice*. CRC press, 2007.

[4] S. Gustafsson, R. Martin, P. Jax, and P. Vary, "A psychoacoustic approach to combined acoustic echo cancellation and noise reduction," *IEEE Transactions on Speech and Audio Processing*, vol. 10, no. 5, pp. 245–256, 2002.

[5] M. M. Halimeh, T. Haubner, A. Briegleb, A. Schmidt, and W. Kellermann, "Combining adaptive filtering and complex-valued deep postfiltering for acoustic echo cancellation," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 121–125, IEEE, 2021.

[6] A. R. Yuliani, M. F. Amri, E. Suryawati, A. Ramdan, and H. F. Pardede, "Speech enhancement using deep learning methods: A review," *Jurnal Elektronika dan Telekomunikasi*, vol. 21, no. 1, pp. 19–26, 2021.

[7] Y. Hu, Y. Liu, S. Lv, M. Xing, S. Zhang, Y. Fu, J. Wu, B. Zhang, and L. Xie, "Dccrn: Deep complex convolution recurrent network for phase-aware speech enhancement," in

Proceedings of Interspeech 2020, (Shanghai, China), ISCA, October 2020. Interspeech 2020, October 25–29, 2020.

[8] R. Peng, L. Cheng, C. Zheng, and X. Li, "Icassp 2021 acoustic echo cancellation challenge: Integrated adaptive echo cancellation with time alignment and deep learning-based residual echo plus noise suppression," in *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 146–150, 2021.

[9] F. Weninger, H. Erdogan, S. Watanabe, E. Vincent, J. Le Roux, J. R. Hershey, and B. Schuller, "Speech enhancement with lstm recurrent neural networks and its application to noise-robust asr," in *International conference on latent variable analysis and signal separation*, pp. 91–99, Springer, 2015.

[10] K. Tan and D. Wang, "A convolutional recurrent neural network for real-time speech enhancement.," in *Interspeech*, vol. 2018, pp. 3229–3233, 2018.

[11] H. Purwins, B. Li, T. Virtanen, J. Schlüter, S.-Y. Chang, and T. Sainath, "Deep learning for audio signal processing," *IEEE Journal of Selected Topics in Signal Processing*, vol. 13, no. 2, pp. 206–219, 2019.

[12] K. Sridhar, R. Cutler, A. Saabas, T. Parnamaa, M. Loide, H. Gamper, S. Braun, R. Aichner, and S. Srinivasan, "Icassp 2021 acoustic echo cancellation challenge: Datasets, testing framework, and results," in *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 151–155, 2021.

[13] J.-M. Valin, S. Tenneti, K. Helwani, U. Isik, and A. Krishnaswamy, "Low-complexity, real-time joint neural echo control and speech enhancement based on percepnet," in *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 7133–7137, 2021.

[14] Z. Wang, Y. Na, Z. Liu, B. Tian, and Q. Fu, "Weighted recursive least square filter and neural network based residual echo suppression for the aec-challenge," in *ICASSP*

2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 141–145, IEEE, 2021.

[15] S. Zhang, Z. Wang, J. Sun, Y. Fu, B. Tian, Q. Fu, and L. Xie, "Multi-task deep residual echo suppression with echo-aware loss," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 9127–9131, IEEE, 2022.

[16] C. K. Reddy, V. Gopal, and R. Cutler, "Dnsmos: A non-intrusive perceptual objective speech quality metric to evaluate noise suppressors," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6493–6497, IEEE, 2021.

[17] M. Purin, S. Sootla, M. Sponza, A. Saabas, and R. Cutler, "Aecmos: A speech quality assessment metric for echo impairment," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 901–905, IEEE, 2022.

[18] J. Benesty, T. Gänsler, D. R. Morgan, M. M. Sondhi, and S. L. Gay, eds., *Advances in Network and Acoustic Echo Cancellation*. Berlin, Germany: Springer, 2001.

[19] A. Oppenheim and R. W. Schafer, *Discrete -time signal processing*. Pearson Higher Education, 2010.

[20] T. Gansler, M. Hansson, C.-J. Ivarsson, and G. Salomonsson, "A double-talk detector based on coherence," *IEEE Transactions on Communications*, vol. 44, no. 11, pp. 1421–1427, 1996.

[21] D. S. Williamson, Y. Wang, and D. Wang, "Complex ratio masking for monaural speech separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 3, pp. 483–492, 2016.

[22] J. P. Borrallo and M. G. Otero, "On the implementation of a partitioned block frequency domain adaptive filter (pbfdaf) for long acoustic echo cancellation," *Signal Processing*, vol. 27, no. 3, pp. 301–315, 1992.

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

[24] C. Avendano, "Acoustic echo suppression in the stft domain," in *Proceedings of the 2001 IEEE Workshop on the Applications of Signal Processing to Audio and Acoustics (Cat. No.01TH8575)*, pp. 175–178, 2001.

[25] P. Scalart *et al.*, "Speech enhancement based on a priori signal to noise estimation," in *1996 IEEE international conference on acoustics, speech, and signal processing conference proceedings*, vol. 2, pp. 629–632, IEEE, 1996.

[26] R. Martin, "Noise power spectral density estimation based on optimal smoothing and minimum statistics," *IEEE Transactions on speech and audio processing*, vol. 9, no. 5, pp. 504–512, 2001.

[27] S. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 27, no. 2, pp. 113–120, 1979.

[28] J. Benesty, S. Makino, and J. Chen, *Speech enhancement*. Springer Science & Business Media, 2006.

[29] D. Wang and J. Chen, "Supervised speech separation based on deep learning: An overview," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 10, pp. 1702–1726, 2018.

[30] Y. Ephraim, "Statistical-model-based speech enhancement systems," *Proceedings of the IEEE*, vol. 80, no. 10, pp. 1526–1555, 1992.

[31] S. Rangachari and P. C. Loizou, "A noise-estimation algorithm for highly non-stationary environments," *Speech communication*, vol. 48, no. 2, pp. 220–231, 2006.

[32] Y. Jiang, B. Tian, H. Wang, S. Zhao, B. Ma, D. Chen, and X. Li, "E2e-aec: Implementing an end-to-end neural network learning approach for acoustic echo cancellation," *arXiv preprint arXiv:2601.16774*, 2026.

[33] Z. Ouyang, H. Yu, W.-P. Zhu, and B. Champagne, "A fully convolutional neural network for complex spectrogram processing in speech enhancement," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5756–5760, IEEE, 2019.

[34] S. R. Park and J. W. Lee, "A fully convolutional neural network for speech enhancement," *Interspeech 2017*, pp. 1993–1997, 08 2017.

[35] O. Abdel-Hamid, A.-r. Mohamed, H. Jiang, L. Deng, G. Penn, and D. Yu, "Convolutional neural networks for speech recognition," *IEEE/ACM Transactions on audio, speech, and language processing*, vol. 22, no. 10, pp. 1533–1545, 2014.

[36] S. Pascual, A. Bonafonte, and J. Serrà, "Segan: Speech enhancement generative adversarial network," *Interspeech 2017*, p. 3642, 2017.

[37] M. N. Ali, A. Brutti, and D. Falavigna, "Speech enhancement using dilated wave-u-net: an experimental analysis," in *2020 27th Conference of Open Innovations Association (FRUCT)*, pp. 3–9, 2020.

[38] A. Graves, A.-r. Mohamed, and G. Hinton, "Speech recognition with deep recurrent neural networks," in *2013 IEEE international conference on acoustics, speech and signal processing*, pp. 6645–6649, IEEE, 2013.

[39] J. Picone, "Fundamentals of speech recognition: A short course," *Institute for Signal and Information Processing, Mississippi State University*, 1996.

[40] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[41] F. Weninger, F. Eyben, and B. Schuller, "Single-channel speech separation with memory-enhanced recurrent neural networks," in *2014 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pp. 3709–3713, IEEE, 2014.

[42] S. Zhang, Y. Kong, S. Lv, Y. Hu, and L. Xie, "Ft-lstm based complex network for joint acoustic echo cancellation and speech enhancement," *arXiv preprint arXiv:2106.07577*, 2021.

[43] S.-W. Fu, Y. Tsao, X. Lu, *et al.*, "Snr-aware convolutional neural network modeling for speech enhancement.," in *Interspeech*, pp. 3768–3772, 2016.

[44] C. Breining, P. Dreiscitel, E. Hansler, A. Mader, B. Nitsch, H. Puder, T. Schertler, G. Schmidt, and J. Tilp, "Acoustic echo control. an application of very-high-order adaptive filters," *IEEE Signal Processing Magazine*, vol. 16, no. 4, pp. 42–69, 1999.

[45] J. Benesty, C. Paleologu, T. Gänsler, and S. Ciochină, "Recursive least-squares algorithms," in *A perspective on stereophonic acoustic echo cancellation*, pp. 63–69, Springer, 2011.

[46] P. Estermann and A. Kaelin, "A hands-free phone system based on a partitioned frequency-domain adaptive echo canceler," in *1996 8th European Signal Processing Conference (EUSIPCO 1996)*, pp. 1–4, IEEE, 1996.

[47] R. Martin, "Spectral subtraction based on minimum statistics," *Proc. EUSIPCO-94, Edinburgh, UK*, 1994.

[48] G. Carbajal, R. Serizel, E. Vincent, and Humbert, "Multiple-input neural network-based residual echo suppression," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 231–235, 2018.

[49] A. Schwarz, C. Hofmann, and W. Kellermann, "Spectral feature-based nonlinear residual echo suppression," in *2013 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pp. 1–4, 2013.

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

[51] Y. Wang and D. Wang, "Towards scaling up classification-based speech separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 7, pp. 1381–1390, 2013.

[52] A. Narayanan and D. Wang, "Ideal ratio mask estimation using deep neural networks for robust speech recognition," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 7092–7096, 2013.

[53] K. Tan and D. Wang, "Learning complex spectral mapping with gated convolutional recurrent networks for monaural speech enhancement," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 380–390, 2020.

[54] Y. Xu, J. Du, L.-R. Dai, and C.-H. Lee, "A regression approach to speech enhancement based on deep neural networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 1, pp. 7–19, 2015.

[55] S.-W. Fu, Y. Tsao, X. Lu, and H. Kawai, "Raw waveform-based speech enhancement by fully convolutional networks," in *2017 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pp. 006–012, 2017.

[56] A. Jansson, E. Humphrey, N. Montecchio, R. Bittner, A. Kumar, and T. Weyde, "Singing voice separation with deep u-net convolutional networks," in *Proceedings of the 18th International Society for Music Information Retrieval Conference (ISMIR)*, (Suzhou, China), ISMIR, October 2017. ISMIR 2017, October 23–27, 2017.

[57] H.-S. Choi, J.-H. Kim, J. Huh, A. Kim, J.-W. Ha, and K. Lee, "Phase-aware speech enhancement with deep complex u-net," in *International Conference on Learning Representations*, 2018.

[58] C. Knapp and G. Carter, "The generalized correlation method for estimation of time delay," *IEEE transactions on acoustics, speech, and signal processing*, vol. 24, no. 4, pp. 320–327, 1976.

[59] B. Farhang-Boroujeny, "Analysis and efficient implementation of partitioned block lms adaptive filters," *IEEE transactions on signal processing*, vol. 44, no. 11, pp. 2865–2868, 1996.

[60] Y. Takahashi, H. Saruwatari, K. Shikano, and K. Kondo, "Musical-noise analysis in methods of integrating microphone array and spectral subtraction based on higher-order statistics," *EURASIP Journal on Advances in Signal Processing*, vol. 2010, no. 1, p. 431347, 2010.

[61] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*, pp. 234–241, Springer, 2015.

[62] Y. N. Dauphin, A. Fan, M. Auli, and D. Grangier, "Language modeling with gated convolutional networks," in *International conference on machine learning*, pp. 933–941, PMLR, 2017.

[63] K. Tan, J. Chen, and D. Wang, "Gated residual networks with dilated convolutions for monaural speech enhancement," *IEEE/ACM transactions on audio, speech, and language processing*, vol. 27, no. 1, pp. 189–198, 2018.

[64] S. Zhang, Z. Wang, Y. Ju, Y. Fu, Y. Na, Q. Fu, and L. Xie, "Personalized acoustic echo cancellation for full-duplex communications," in *Proc. Interspeech 2022*, pp. 2518–2522, 2022.

[65] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *International conference on machine learning*, pp. 448–456, pmlr, 2015.

[66] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," in *Proceedings of the IEEE international conference on computer vision*, pp. 1026–1034, 2015.

[67] D. Stoller, S. Ewert, and S. Dixon, "Wave-u-net: A multi-scale neural network for end-to-end audio source separation," *arXiv preprint arXiv:1806.03185*, 2018.

[68] J. Li, A. Mohamed, G. Zweig, and Y. Gong, "Lstm time and frequency recurrence for automatic speech recognition," in *2015 IEEE workshop on automatic speech recognition and understanding (ASRU)*, pp. 187–191, IEEE, 2015.

[69] S. Gannot, E. Vincent, S. Markovich-Golan, and A. Ozerov, "A consolidated perspective on multimicrophone speech enhancement and source separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 4, pp. 692–730, 2017.

[70] A. Katrompas and V. Metsis, "Enhancing lstm models with self-attention and stateful training," in *Proceedings of SAI Intelligent Systems Conference*, pp. 217–235, Springer, 2021.

[71] A. Odena, V. Dumoulin, and C. Olah, "Deconvolution and checkerboard artifacts," *Distill*, vol. 1, no. 10, p. e3, 2016.

[72] Y. Luo, Z. Chen, J. R. Hershey, J. Le Roux, and N. Mesgarani, "Deep clustering and conventional networks for music separation: Stronger together," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 61–65, 2017.

[73] S.-W. Fu, C.-F. Liao, Y. Tsao, and S.-D. Lin, "Metricgan: Generative adversarial networks based black-box metric scores optimization for speech enhancement," in *International Conference on Machine Learning*, pp. 2031–2041, PmLR, 2019.

[74] D. Yin, C. Luo, Z. Xiong, and W. Zeng, "Phasen: A phase-and-harmonics-aware speech enhancement network," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 34, pp. 9458–9465, 2020.

[75] Y. Xia, S. Braun, C. K. A. Reddy, H. Dubey, R. Cutler, and I. Tashev, "Weighted speech distortion losses for neural-network-based real-time speech enhancement," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 871–875, 2020.

[76] J. M. Martin-Doñas, A. M. Gomez, J. A. Gonzalez, and A. M. Peinado, "A deep learning loss function based on the perceptual evaluation of the speech quality," *IEEE Signal Processing Letters*, vol. 25, no. 11, pp. 1680–1684, 2018.

[77] S. E. Eskimez, X. Wang, M. Tang, H. Yang, Z. Zhu, Z. Chen, H. Wang, and T. Yoshioka, "Human listening and live captioning: Multi-task training for speech enhancement," in *Proc. Interspeech 2021*, pp. 2686–2690, 2021.

[78] J.-M. Valin, U. Isik, N. Phansalkar, R. Giri, K. Helwani, and A. Krishnaswamy, "A perceptually-motivated approach for low-complexity, real-time enhancement of fullband speech," in *Proceedings of Interspeech 2020*, ISCA, October 2020. Interspeech 2020, October 25–29, 2020.

[79] J. L. Roux, S. Wisdom, H. Erdogan, and J. R. Hershey, "Sdr – half-baked or well done?," in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 626–630, 2019.

[80] G. Enzner, R. Martin, and P. Vary, "Unbiased residual echo power estimation for hands-free telephony," in *2002 IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 2, pp. II–1893–II–1896, 2002.

[81] C. K. A. Reddy, H. Dubey, V. Gopal, R. Cutler, S. Braun, H. Gamper, R. Aichner, and S. Srinivasan, "Icassp 2021 deep noise suppression challenge," in *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6623–6627, 2021.

[82] B. Naderi and R. Cutler, "An open source implementation of itu-t recommendation p.808 with validation," in *Proceedings of Interspeech 2020*, ISCA, International Speech Communication Association (ISCA), October 2020. Interspeech 2020, October 25–29, 2020.

[83] J. F. Gemmeke, D. P. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter, "Audio set: An ontology and human-labeled dataset for audio events," in *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pp. 776–780, IEEE, 2017.

[84] F. Font, G. Roma, and X. Serra, "Freesound technical demo," in *Proceedings of the 21st ACM international conference on Multimedia*, pp. 411–412, 2013.

[85] J. Thiemann, N. Ito, and E. Vincent, "The diverse environments multi-channel acoustic noise database (demand): A database of multichannel environmental noise recordings," in *Proceedings of Meetings on Acoustics*, vol. 19, p. 035081, Acoustical Society of America, 2013.

[86] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *International Conference on Learning Representations*, 2019.

[87] H. Schröter, A. Maier, A. Escalante-B, and T. Rosenkranz, "Deepfilternet2: Towards real-time speech enhancement on embedded devices for full-band audio," in *2022 International Workshop on Acoustic Signal Enhancement (IWAENC)*, pp. 1–5, 2022.

[88] R. Pascanu, T. Mikolov, and Y. Bengio, "On the difficulty of training recurrent neural networks," in *Proceedings of the 30th International Conference on International Conference on Machine Learning-Volume 28*, pp. III–1310, 2013.

[89] H. Il Koh, S. Na, and M. N. Kim, "Lsenet: A lightweight spectral enhancement network for high-quality speech processing on resource-constrained platforms," *IEEE Access*, vol. 13, pp. 116934–116943, 2025.

[90] P. Goyal, P. Dollár, R. Girshick, P. Noordhuis, L. Wesolowski, A. Kyrola, A. Tulloch, Y. Jia, and K. He, "Accurate, large minibatch sgd: Training imagenet in 1 hour," *arXiv preprint arXiv:1706.02677*, 2017.

[91] A. S. Meyer, "Timing in conversation," *Journal of Cognition*, vol. 6, no. 1, p. 20, 2023.

[92] D. Stefani, S. Peroni, L. Turchet, *et al.*, "A comparison of deep learning inference engines for embedded real-time audio classification," in *Proceedings of the International Conference on Digital Audio Effects, DAFx*, vol. 3, pp. 256–263, DAFx, 2022.

[93] I. Fedorov, M. Stamenovic, C. Jensen, L.-C. Yang, A. Mandell, Y. Gan, M. Mattina, and P. N. Whatmough, "Tinylstms: Efficient neural speech enhancement for hearing aids," *arXiv preprint arXiv:2005.11138*, 2020.

[94] E. Seidel, G. Enzner, P. Mowlaee, and T. Fingscheidt, "Neural kalman filters for acoustic echo cancellation: Comparison of deep neural network-based extensions," *IEEE Signal Processing Magazine*, vol. 41, no. 6, pp. 24–38, 2025.

[95] L. Cheng, R. Peng, A. Li, C. Zheng, and X. Li, "Deep learning-based stereophonic acoustic echo suppression without decorrelation," *The Journal of the Acoustical Society of America*, vol. 150, no. 2, pp. 816–829, 2021.