Advances in Intelligent Robotics and AI Systems

Academixs Open Access Publishers

Research Article

Multi-Input Deep Complex Convolution Recurrent Network for Joint Acoustic Echo Cancellation and Background Noise Suppression in Real-Time Speech Communication

Kum-Song Pak[1], Chol-I Om[3], Hyon U Kong[2,4*], Kwon Kim[1], Chol-Ui Ri[1] and Chol-Nam Om[1]

[1]*Institute of Information Technology, Hightech Research and Development Centre, Kim Il Sung University, Pyongyang, Democratic People's Republic of Korea*

[2]*School of Artificial Intelligence, Kim Il Sung University, Pyongyang, Democratic People's Republic of Korea*

[3]*Faculty of Chemistry, University of Sciences, Pyongyang, Democratic People's Republic of Korea*

[4]*School of Business Administration, Northeastern University, Shenyang, 110819, China*

***Corresponding Author:**
Hyon U Kong. School of Business Administration, Northeasten University, Shenyang, 110819, China.

Received Date: 08.07.2026 **Accepted Date:** 17.07.2026 **Published Date:** 25.07.2026

Abstract

The proliferation of online communication platforms has intensified demand for high-quality speech enhancement in adverse acoustic environments. This paper presents a Multi-Input Deep Complex Convolution Recurrent Network (MIDCCRN) that extends conventional DCCRN architectures to handle multiple input streams through temporal concatenation, enabling unified processing of microphone signals, far-end references, and intermediate estimates. We propose a computationally efficient Complex LSTM implementation utilizing complex convolution operations, which reduces activation function calls by 50% while maintaining equivalent representational capacity. Building upon these innovations, we introduce a three-stage cascaded framework for joint acoustic echo cancellation (AEC) and background noise suppression (BNS) comprising echo estimation, acoustic echo cancellation, and background noise suppression modules, trained end-to-end with multi-task learning and a novel short-time weighted signal-to-distortion ratio (STwSDR) loss function. Comprehensive experiments on the ICASSP 2022 AEC Challenge dataset demonstrate state-of-the-art performance with a final score M = 0.85, comparable to the challenge-winning solution, while maintaining real-time capability (RTF = 0.302) and modest model size (4.28M parameters). Ablation studies validate the effectiveness of each proposed component.

Keywords: Acoustic Echo Cancellation, Background Noise Suppression, Multi-Input Deep Complex Convolution Recurrent Network, Speech Enhancement, Real-Time Communication, Complex Neural Networks

Introduction

Background and Motivation

The rapid globalization of information and communication technology has fundamentally transformed human interaction patterns with multimedia communication systems becoming indispensable infrastructure for remote collaboration, telemedicine, and distance education [1,2]. The COVID-19 pandemic served as a catalyst for widespread adoption of online communication tools embedding video conferencing into daily professional and social activities [3]. However, audio quality in these systems frequently degrades due to acoustic echo, background noise, reverberation, and interfering speech significantly compromising speech intelligibility and user experience [4,5].

Among these impairments, acoustic echo cancellation (AEC) and background noise suppression (BNS) constitute the most critical challenges for full-duplex communication systems. Acoustic echo arises when far-end speech played through local loudspeakers is captured by the near-end microphone, creating distracting feedback loops. Background noise, ranging from stationary mechanical hum to non-stationary human activity and environmental sounds, masks target speech and reduces intelligibility. The simultaneous presence of both impairments

during double-talk periods—when near-end and far-end speakers are active concurrently—poses particularly severe difficulties for conventional signal processing methods. Traditional digital signal processing (DSP) approaches for BNS primarily rely on spectral suppression techniques applied in the time-frequency domain, including Wiener filtering, spectral subtraction, and statistical model-based methods. While these methods can be effective for stationary noise, they suffer from the fundamental trade-off between noise attenuation and speech distortion, particularly under low signal-to-noise ratio conditions with non-stationary interference. Multi-channel beamforming techniques leverage spatial information from microphone arrays to achieve superior noise reduction without excessive distortion but their applicability is limited by hardware constraints in consumer devices [6,7].

For AEC, adaptive filtering algorithms—such as normalized least mean squares (NLMS), recursive least squares (RLS), and affine projection algorithms—have been extensively studied [8,9]. However, these methods exhibit slow convergence in dynamic acoustic environments, performance degradation under double-talk conditions, and inability to model nonlinear distortions introduced by low-cost loudspeakers and power amplifiers [10,11].

The emergence of deep neural networks (DNNs) has enabled paradigm shifts in both BNS and AEC [12,13]. Deep learning approaches can learn complex spectral mappings from large datasets, offering superior generalization to unseen acoustic conditions compared to model-based methods. Convolutional recurrent networks (CRNs) have proven particularly effective, combining the local pattern extraction capabilities of convolutional neural networks with the temporal modeling capacity of recurrent neural networks [14,15]. The Deep Complex Convolution Recurrent Network (DCCRN) further advanced this direction by operating directly on complex spectrograms, jointly optimizing magnitude and phase responses through complex-valued convolutions and complex LSTM (CLSTM) layers.

Despite these advances, three critical limitations persist in current approaches:
First, existing complex networks predominantly process single-channel inputs, limiting their applicability to AEC scenarios where both microphone signals and far-end references must be jointly processed. While simple input concatenation has been explored, principled integration of multiple signals within complex-valued architectures remains underdeveloped.

Second, computational inefficiency plagues current implementations. Complex LSTM implementations in existing frameworks require redundant computations, with separate real and imaginary pathway processing that doubles activation function evaluations. This inefficiency hinders deployment in resource-constrained real-time systems.

Third, end-to-end training of joint AEC-BNS systems often struggles with the fundamental tension between aggressive echo suppression and speech preservation. Stage-wise processing—explicitly modeling echo paths before noise removal—has demonstrated advantages, but integration with complex-valued deep networks remains unexplored.

Related Work

Deep Learning for Background Noise Suppression

Recent BNS methods utilizing DNNs can be categorized into regression-based and mask-based approaches [16,17]. Regression methods directly predict clean speech spectrograms or waveforms from noisy inputs, while mask-based methods estimate time-frequency masks applied to the mixture. The choice of target representation significantly impacts performance: ideal ratio masks provide stable training targets, while complex ratio masks preserve phase information critical for high-fidelity reconstruction [18,19].

CRN architectures have emerged as a dominant paradigm, with encoder-decoder structures extracting hierarchical features and recurrent bottleneck layers capturing long-range temporal dependencies [20,21]. The DCCRN architecture extended this framework to complex-valued processing, demonstrating that complex convolutions and batch normalization better model phase relationships than real-valued alternatives [22,23]. However, DCCRN and its variants remain limited to single-input scenarios.

Recent work has explored efficiency improvements. Lightweight designs employing depthwise separable convolutions attention mechanisms and state-space models have achieved competitive performance with reduced computational footprints. However, these approaches generally operate on magnitude spectrograms or real-valued features, sacrificing phase modeling capabilities [24-27].

Deep Learning for Acoustic Echo Cancellation

Deep learning approaches for AEC have evolved through several stages. Initial efforts replaced adaptive filters with fully-connected or recurrent networks mapping far-end and microphone signals to near-end estimates. However, these methods struggled with generalization to unseen room acoustics and nonlinear characteristics. The ICASSP AEC Challenge series, initiated in 2021 in response to pandemic-driven demand for robust teleconferencing catalyzed rapid innovation. Winning entries typically combined traditional adaptive filtering for linear echo cancellation with neural networks for residual echo suppression and noise removal [27-29]. Multi-stage cascaded architectures have proven particularly effective, with explicit echo estimation preceding joint suppression.

Recent advances include the F-T-LSTM based complex network which combines frequency-time domain processing with complex-valued representations, achieving superior performance to real-valued alternatives [30]. Task-specific architectures such as SMRU (Li et al., 2024) introduce configurable complexity through variable frame rate processing, while CAGCRN (Wang et al., 2025) achieves extreme lightweight design (0.07M parameters) through cross-attention gating. However, these methods typically process concatenated inputs in early layers rather than maintaining multi-stream processing throughout the network [31,32].

Joint AEC and BNS

Joint processing of echo cancellation and noise suppression presents unique challenges, as aggressive noise removal may distort residual echo characteristics, while echo suppression can amplify noise artifacts. Recent approaches address this through

multi-task learning, cascaded specialization or unified estimation with sophisticated loss functions [33]. Wang, Wichern, and Le Roux (2021) demonstrated that leveraging intermediate low-distortion target estimates in cascaded structures improves overall performance compared to end-to-end training [34]. Braun and Valero (2022) validated that explicit task splitting for echo and noise removal outperforms joint estimation, providing theoretical justification for modular architectures [33]. However, existing cascaded systems predominantly employ real-valued networks, leaving potential gains from complex-valued processing unrealized.

Contributions

To address the identified limitations, this paper introduces the Multi-Input Deep Complex Convolution Recurrent Network (MIDCCRN) and its application to joint AEC and BNS. Our specific contributions are:(1) We propose MIDCCRN, extending DCCRN to handle arbitrary K auxiliary inputs through temporal concatenation in the complex spectrogram domain. This design enables unified processing of microphone signals, far-end references, and intermediate estimates while preserving frequency resolution and phase relationships. (2) We implement CLSTM using complex convolution operations rather than separate real/imaginary pathway processing, reducing activation function calls by 50% while maintaining equivalent representational capacity. Comprehensive ablation studies validate that this modification preserves enhancement quality. (3) We design an explicit three-stage architecture for joint AEC-BNS: echo estimation, acoustic echo cancellation, background noise suppression.

Each stage employs task-specific MIDCCRN configurations, with multi-task training using our proposed short-time weighted SDR (STwSDR) loss function. (4) We conduct extensive experiments on the ICASSP 2022 AEC Challenge dataset, including ablation studies validating architectural choices and comparisons with state-of-the-art methods from 2020–2025. Results demonstrate superior performance-complexity trade-offs suitable for real-time deployment. The remainder of this paper is organized as follows: Section 2 details the MIDCCRN architecture, including the multi-input transformation, encoder-decoder structure, and efficient CLSTM implementation. Section 3 presents the three-stage joint AEC-BNS framework, training methodology, and loss functions. Section 4 describes experimental setup, ablation studies, and comparative results. Section 5 concludes with directions for future work.

Multi-Input Deep Complex Convolution Recurrent Network

This section introduces the MIDCCRN architecture. We first describe the input transformation enabling multi-signal processing, then detail the encoder-decoder structure with efficient CLSTM implementation, and finally discuss online deployment considerations.

Figure 1 shows the structure of the MIDCCRN. MIDCCRN consists of three blocks; input signal transform block, encoder-decoder block, and output block.

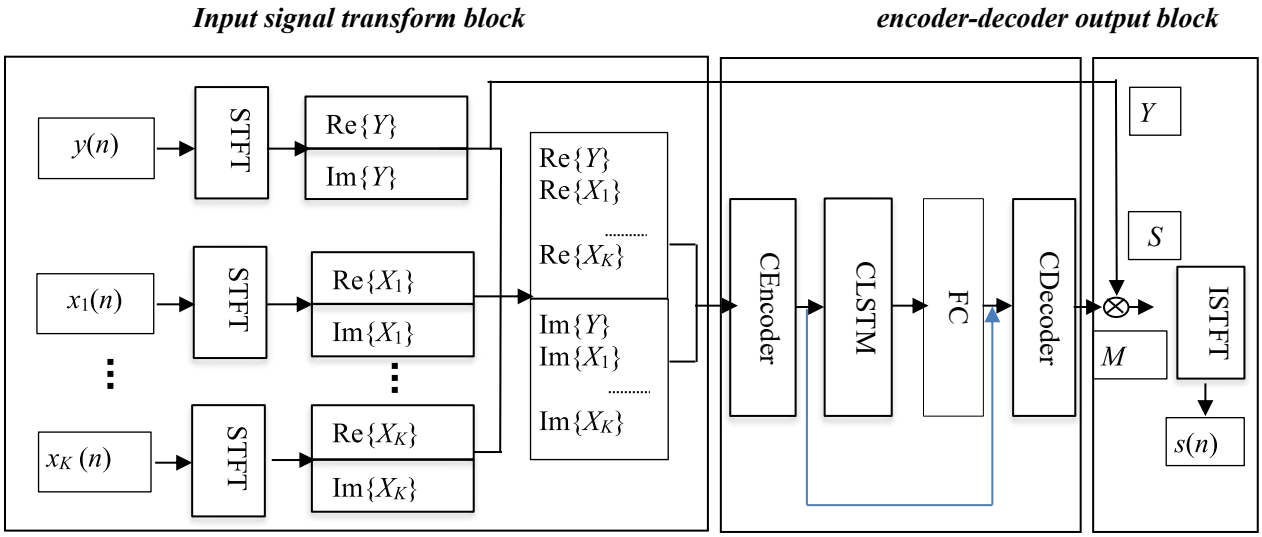

Figure 1: structure of the MIDCCRN

Input signal transform block

Conventional DCCRN processes single-channel complex spectrograms. To enable multi-input scenarios required for AEC, we extend this to accept K auxiliary signals alongside the primary microphone signal. Let $y(n)$ denote the microphone signal and $x_k(n)$ for k=1,...,K denote auxiliary reference signals (e.g., far-end speech). Their short-time Fourier transforms (STFT) yield complex spectrograms $Y(t,f)$ and $X_k(t,f)$, where t and f index time and frequency bins, respectively. The input transformation rearranges these representations by separating real and imaginary components:

$$R_o = [\mathrm{Re}\{Y\}, \mathrm{Re}\{X_1\}, \cdots, \mathrm{Re}\{X_K\}] \in \mathbb{R}^{T\times F\times(K+1)},$$
$$I_o = [\mathrm{Im}\{Y\}, \mathrm{Im}\{X_1\}, \cdots, \mathrm{Im}\{X_K\}] \in \mathbb{R}^{T\times F\times(K+1)} \quad (1)$$

where T and F denote temporal and frequency dimensions. The complex input tensor is then formed as: $U=R_o+ j \cdot I_o \in \mathbb{C}^{T\times F\times(K+1)}$

We employ temporal concatenation along the feature dimension rather than channel-wise stacking. This design preserves the frequency resolution F=161 (for 16 kHz sampling with

512-point FFT and 25ms window), enabling the encoder to learn frequency-dependent phase relationships across input streams. Channel-wise concatenation would require modified encoder architectures and might obscure cross-signal temporal alignment patterns crucial for echo cancellation.

We denote MIDCCRN with K auxiliary inputs as MIDCCRN(K). MIDCCRN (0) corresponds to the original single-input DCCRN configuration. All signals are processed with consistent parameters: sampling rate 16 kHz, Hamming window length 25 ms, hop size 6.25 ms, FFT length 512. Due to conjugate symmetry, only 161 frequency bins are retained. To capture temporal context, 16 consecutive frames are processed jointly, yielding input dimensions T=16(K+1) , F=161 .

Encoder-Decoder Block

The core processing follows a U-Net encoder-decoder architecture with complex-valued operations and skip connections (Fig. 2). Table 1 summarizes layer configurations.

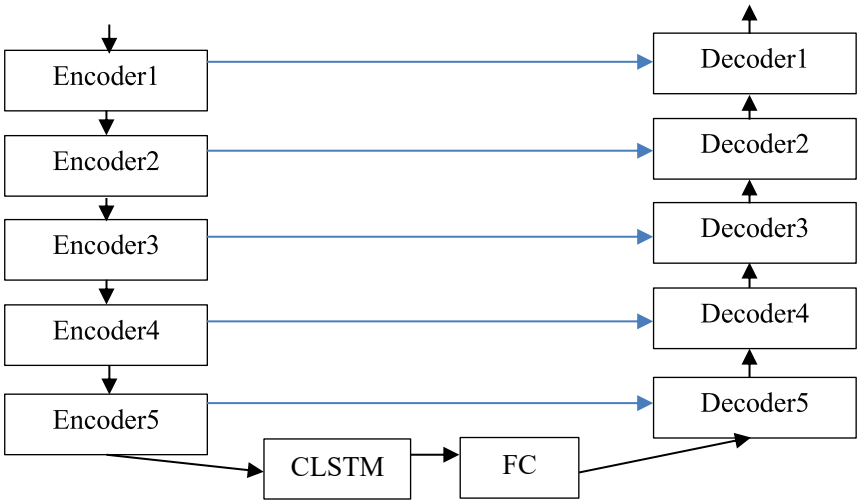

Figure 2: Encoder-Decoder block

layer name	input size	hyper-parameters	output size
Encoder1	2 × T × 161	1 × 3, (1, 2), 16	16 ×T × 80
Encoder2	16 × T × 80	1 × 3, (1, 2), 32	32 ×T × 39
Encoder3	32 × T × 39	1 × 3, (1, 2), 64	64 ×T × 19
Encoder4	64 ×T × 19	1 × 3, (1, 2), 128	128 ×T × 9
Encoder5	128 × T × 9	1 × 3, (1, 2), 256	256 ×T × 4
reshape 1	256 × T × 4		T × 1024
CLSTM	T × 1024	1024	T × 1024
FC	T × 1024	1024	T × 1024
reshape 2	T × 1024		256 ×T × 4
Decoder5 (× 2)	512 ×T × 4	1 × 3, (1, 2), 128	128 ×T × 9
Decoder4 (× 2)	256 ×T × 9	1 × 3, (1, 2), 64	64 ×T × 19
Decoder3 (× 2)	128 ×T × 19	1 × 3, (1, 2), 32	32 ×T × 39
Decoder2 (× 2)	64 ×T × 39	1 × 3, (1, 2), 16	16 ×T × 80
Decoder1(× 2)	32 ×T × 80	1 × 3, (1, 2), 1	1 ×T × 161
concat	1× T ×161 (× 2)		2 ×T × 161

Table 1: Hyper-parameters of Encoder-Decoder block

Table 1 shows the hyper parameters of the encoder-decoder block. In the table, T = 16(M+1) is the time dimension, and the frequency dimension F decreases by half from 161 to 9 in the encode, then increases again in the decoder to 161. The number of channels increases to 2, 16, 32, 64, 128, and 256 in the encoder, and then decreases again to 2 in the decoder. The input size and output size of each layer are given in the form of (number of channels × time dimension × frequency dimension). The hyper-parameters of the layers are given in the form of (kernel size, step size, and number of channels).

Complex Encoder

Each encoder layer comprises complex 2D convolution (CConv2d), complex batch normalization (CBN), and parametric ReLU (PReLU) activation.

For complex filter $W = W_r + jW_i$ and input $X = X_r + jX_i$, the operation is:

$$Y = X \circledast W = (X_r * W_r - X_i * W_i) + j(X_r * W_i + X_i * W_r) \quad (2)$$

implemented via four real-valued convolutions.

Normalizes complex values by computing mean and variance over both real and imaginary components [38]. PReLU is Applied separately to real and imaginary parts: $PReLU(x) = max(0,x)+\alpha\, min(0,x)$, with learnable parameter α .

Efficient Complex LSTM Implementation

The original DCCRN implements CLSTM using four separate LSTMs processing real and imaginary components independently [14]:

$$\mathbf{Y}_{rr} = \mathrm{LSTM}(\mathbf{X}_r, \mathbf{W}_r),\ \mathbf{Y}_{ir} = \mathrm{LSTM}(\mathbf{X}_i, \mathbf{W}_r),\ \mathbf{Y}_{ri} = \mathrm{LSTM}(\mathbf{X}_r, \mathbf{W}_i),\ \mathbf{Y}_{ii} = \mathrm{LSTM}(\mathbf{X}_i, \mathbf{W}_i) \quad (3)$$

$$\mathbf{Y} = (\mathbf{Y}_{rr} - \mathbf{Y}_{ii}) + \boldsymbol{j}(\mathbf{Y}_{ri} + \mathbf{Y}_{ir})$$

This requires 12 activation function evaluations per time step (3 gates × 4 pathways).

We replace real-valued convolutions within LSTM gates with complex convolutions, enabling single-pathway processing:

$$\mathbf{i}_t = \boldsymbol{\sigma}(\mathbf{W}_{xi} * \mathbf{x}_t + \mathbf{W}_{hi} * \mathbf{h}_{t-1} + \mathbf{b}_t) \quad (4)$$

$$\mathbf{f}_t = \boldsymbol{\sigma}(\mathbf{W}_{xf} * \mathbf{x}_t + \mathbf{W}_{hf} * \mathbf{h}_{t-1} + \mathbf{b}_f) \quad (5)$$

$$\mathbf{c}_t = \mathbf{f}_t \circ \mathbf{c}_{t-1} + \mathbf{i}_t \circ \boldsymbol{tanh}(\mathbf{W}_{xc} * \mathbf{x}_t + \mathbf{W}_{hc} * \mathbf{h}_{t-1} + \mathbf{b}_c) \quad (6)$$

$$\mathbf{o}_t = \boldsymbol{\sigma}(\mathbf{W}_{xo} * \mathbf{x}_t + \mathbf{W}_{ho} * \mathbf{h}_{t-1} + \mathbf{b}_o) \quad (7)$$

$$\mathbf{h}_t = \mathbf{o}_t \circ \boldsymbol{tanh}(\mathbf{c}_t) \quad (8)$$

where * denotes complex convolution and ○ complex element-wise multiplication.

The proposed implementation requires only 6 activation calls (3 gates × 1 pathway with real/imaginary pairs), achieving 50% reduction compared to the conventional approach. Complex multiplication requires four real multiplications, but modern accelerators optimize this overhead.

The primary savings come from reduced memory access patterns and activation function evaluations, which are significant on edge devices. Section 4.3.1 validates that this modification preserves enhancement performance while reducing computational cost, addressing concerns that the reduction might merely merge operations without actual savings.

Complex Decoder

Decoder layers employ transposed complex convolutions for up sampling. Skip connections concatenate encoder features with decoder activations, preserving fine-grained details. The final layer outputs a complex ratio mask (CRM) $M \in \mathbb{C}^{T\times F}$.

Output Block

The Encoder-Decoder block outputs a complex rate mask (CRM) M. Multiplying this complex mask M by the complex spectrum Y of the noisy signal yields the spectrum of the clean signal.

$$\hat{S} = M \cdot Y \quad (14)$$

The polar coordinate representation $(M_m\ ,\ M_\varphi)$ of the Cartesian representation of the complex number $M = M_r + jM_i\ (M_r, M_i \in \mathbb{R}, j = \sqrt{-1})$ is as follows;

$$M_m = \sqrt{{M_r}^2 + {M_i}^2},\ M_\varphi = arctan2(M_i, M_r), \quad (15)$$

where $arctan2$ is a quarter-square tangent function with values in the interval $(-\pi,\pi]$. The spectrum of the clean signal is calculated as follows;

$$\hat{S} = Y_m \cdot M_m exp\{Y_\varphi + M_\varphi\}, \quad (16)$$

where $(Y_m,\ Y_\varphi)$ is the polar coordinate representation of the spectrum Y of the incoming noisy signal and $\exp\{\cdot\}$ is an exponential function. The time-domain clean signal $\hat{s} = \mathrm{ISTFT}\{\hat{S}\}$ is estimated.

Online Implementation Considerations

For deployment in real-time communication systems, the following design constraints are explicitly addressed:

The system employs unidirectional CLSTM without future context access. All convolutions are causal (no padding on future frames), ensuring no algorithmic latency beyond the inherent STFT windowing. Total latency comprises the STFT analysis window (25 ms) plus context buffer (100 ms for 16 frames at 6.25 ms hop), totalling 125 ms. This satisfies ITU-T G.114 recommendations for conversational quality (< 150 ms one-way). Frame-by-frame processing maintains O (1) memory complexity per time step through sliding window buffer management. The 16-frame context is maintained in a circular buffer, updating incrementally.

As validated in Section 4, RTF = 0.302 on Intel Xeon E5-2690 CPU, satisfying real-time constraints (RTF < 1.0).

Performance of AEC and BNS

This section presents the three-stage cascaded architecture utilizing MIDCCRN modules for joint acoustic echo cancellation and background noise suppression.

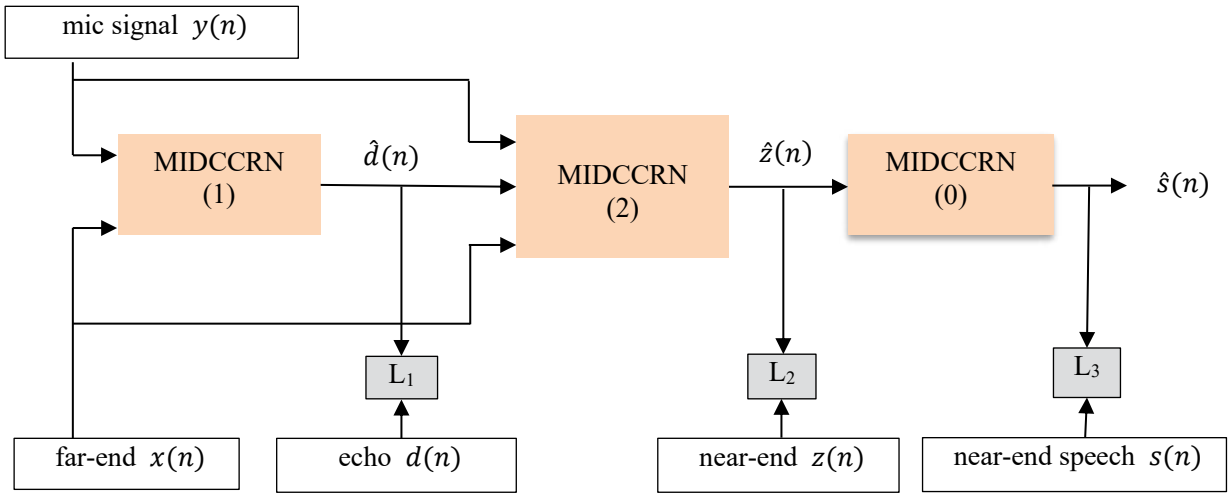

Figure 3: The framework for AEC and BNS

Problem Formulation

Consider the standard AEC signal model:

$$y(n) = d(n) + z(n) = d(n) + s(n) + v(n) \quad (17)$$

where $y(n)$ is the microphone signal, $d(n) = \mathcal{H}\{x(n)\}$ the acoustic echo (far-end $x(n)$ convolved with room impulse response $\mathcal{H}$), s(n) the near-end speech, $v(n)$ the background noise, and $z(n)$=$s(n)$+$v(n)$ the near-end signal. The joint AEC-BNS objective is to estimate clean near-end speech $\hat{s}(n)$ from microphone $y(n)$ and far-end reference $x(n)$, suppressing both echo $d(n)$ and noise $v(n)$ while minimizing distortion of $s(n)$.

Three-Stage Cascaded Architecture

Our framework comprises three sequentially connected MIDCCRN modules (Fig. 3). As shown in the figure, the proposed network consists of three blocks. The first is the echo estimation block, the second is the AEC block, and the third is the BNS block. L1, L2, and L3 are loss functions for echo signal, near-end signal, and near-end signals, respectively. $\hat{s}(n)$ denotes the estimate of $s(n)$. The first block inputs the microphone signal $y(n)$ and the far-end signal $x(n)$ into MDCCRN (1) and estimates the echo signal $\hat{d}(n)$. The second block inputs the microphone signal $y(n)$, the far-end signal $x(n)$, and the estimated echo signal $\hat{d}(n)$ into MDCCRN (2), removes the echo and estimates the near-end signal $\hat{z}(n)$. The third block inputs the estimated near-end signal $\hat{z}(n)$ into MDCCRN (0), removes noise and estimates the target near-end speech signal $\hat{s}(n)$. In the previous work, we usually use the weighted-signal-to-distortion ration (wed) as a loss function. Let $y,s,\hat{s}$ be noisy signal (mixed signal), clean signal, and estimated signal in the time-domain, respectively.

$$\mathrm{wSDR}(y, s, \hat{s}) = -\alpha\varphi(s, \hat{s}) - (1-\alpha)\varphi(n, \hat{n}), \quad (18)$$

where $\varphi(v,w)$ represents the cross-correlation between two vectors v,w, i.e., the cosine of the angle between the two vectors.

$$\varphi(v, w) = \frac{v^T w}{\|v\|\|w\|}, \quad (19)$$

where $\|\cdot\|$ is the Euclidean norm? And $\alpha = \frac{\|s\|^2}{\|y\|^2}$ is the energy ratio of the clean signal to the noisy signal. Also, $n = y - s$ and $\hat{n} = y - \hat{s}$ are the noise and the estimated noise, respectively.

The wed calculates the loss function for the total length of the signal. Applying wed to segment the whole signal into short time intervals can better reflect the local characteristics of the signal. So, we use the short-time-wed (STwSDR) as a loss function.

The sequence of signals segmented by N short time intervals of $y,s,\hat{s}$ can be written as

$$y = \{y_0, \cdots, y_N\},\ s = \{s_0, \cdots, s_N\},\ \hat{s} = \{\hat{s}_0, \cdots, \hat{s}_N\}.$$

Then STwSDR is defined as follows;

$$\mathrm{STwSDR}(y, s, \hat{s}) = \frac{\sum_{i=0}^{N-1} wSDR(y_i, s_i, \hat{s}_i)}{N} \quad (20)$$

The number of segments N is usually set to 10. In Fig. 3, all loss functions L_1, L_2 and L_3 are STwSDR.

We train the model using multi-task learning. The final loss function is as follows;

$$Loss = \beta_1 \cdot \mathrm{STwSDR}_1\left(y(n), d(n), \hat{d}(n)\right) + \beta_2 \cdot \mathrm{STwSDR}_2(y(n), z(n), \hat{z}(n)) + \beta_3 \cdot \mathrm{STwSDR}_3\left(y(n), s(n), \hat{s}(n)\right) \quad (21)$$

Here, β_1, β_2 and β_3 are respectively set to values between 0 and 1 as constants determining the weight of each loss term.

$$\beta_1, \beta_2, \beta_3 \in [0, 1] \quad (22)$$

The network proposed in this section can effectively remove echo and noise.

Experiments and Discussion

In this section, we present a performance evaluation of our proposed method for BNS and AEC through experiments

Experimental Design

Experiments utilize the ICASSP 2022 AEC Challenge synthetic dataset, comprising 10,000 10-second utterances encompassing. The dataset is randomly divided into 80% for training, 10% for

validation and 10% for testing. The length of each signal is 10s. is resampled at 16k Hz. The window length is set as 25ms, the hop size is set as 6.25ms and the FFT length is set as 512. The optimizer is selected as Adam. The initial learning rate is set to 0.001.

As objective evaluation metrics for acoustic echo cancellation, echo return loss enhancement (ERLE), signal-to-artefact-ratio (SAR), and scale-invariant source-to-noise ratio (SI-SNR) are used. The mean opinion score (MOS) is used as a subjective metric. We use an objective perceptual quality assessment scale called AECMOS.

First, we determine the parameters β_1, β_2, and β_3 in the proposed method through simulations. Next, we compare the proposed MIDCCRN with 2021 AEC challenge baseline, F-T-LSTM based complex network, and best methods in the 2022 acoustic echo cancellation challenge contest.

Final score for the AEC challenge performance evaluation is determined using the weighted average of the four subjective scores (M_1, M_2, M_3, M_4) and Wacc.

$$M = \sum_{i=1}^{4} \frac{1}{4} \alpha_i (M_i - 1) + \alpha_5 Wacc \quad (24)$$

where M_1 is far end single talk, M_2 is near end single talk, M_3 is double talk echo, and M_4 is double talk other. The word accuracy rate (Wacc) is used as a metric in the speech recognition and AEC challenge; Wacc = 1 – WER, where WER denotes word error rate. We all the weights are set equally. $\alpha_i = 1/5$. We also use real-time factor (RTF) and network size for performance evaluation.

Experimental Results

Loss Weight Optimization

First, let us determine experimentally the best value of the parameters β_1, β_2, and β_3 in the final loss function formula (21) of our proposed method. The final BNS block MIDCCRN (0) must be and plays the most important role. So, the loss function weight of the last BNS block is always set to 1 and the other parameters are changed.

Table 2 shows the results of SI-SNR comparison for different values of the parameters β_1, β_2, and β_3 of the final loss function for the proposed method. In the table, C is the clean signal, N-F is the noisy far-end signal, N-N is the noisy near-end signal, and N-FN is the noisy signal both the far-end signal and the near-end speech signal. As can be seen from the table, if

$$\beta_1 = 0.2, \beta_2 = 0.5, \beta_3 = 1, \quad (23)$$

SI-SNR is the best. In the experiments with the proposed method, the parameters are set as Eq. (23).

β_1	β_2	β_3	C	N-F	N-N	N-FN
0.2	0.5	1	13.89	10.52	6.35	6.02
0.2	1	1	13.60	10.32	6.12	5.75
1	1	1	13.58	10.14	5.94	5.34
0.2	0	1	13.37	10.25	5.99	5.68
0	1	1	13.41	10.21	5.94	5.51
0	0	1	12.91	9.92	6.01	5.54
1	0	1	12.47	9.36	5.34	4.81

Table 2: SI-SNR evaluation according parameters $\beta_1, \beta_2, \beta_3$

Oobjective Evaluation

Table 3 shows the results of the comparative evaluation of SAR in present of only near-end signal and ERLE in present of far-end signal alone without near-end signal. As can be seen from the table, the highest SAR and ERLE of the proposed method are found.

method	SAR	ERLE
Baseline	14.70	17.74
F-T-LSTM	17.56	23.98
proposed	**17.85**	**24.62**

Table 3: SAR and ERLE evaluation [dB]

method	C	N-F	N-N	N-FN
Unprocessed	7.60	2.97	-3.65	-3.82
Baseline	12.82	8.05	1.89	1.62
F-T-LSTM	13.51	10.35	6.20	5.78
proposed	**13.89**	**10.52**	**6.35**	**6.02**

Table 4: SI-SNR evaluation under several conditions [dB]

method	MOS
Baseline	3.89
F-T-LSTM	4.14
proposed	4.15

Table 5: MOS evaluation

We compare the SI-SNR of each method through experiments. Table 4 shows the comparison of the SI-SNR [dB]. In the method column, 0 represents the case where no processing is done. As can be seen from the table, the proposed method outperforms all other methods in SI-SNR performance under all conditions.

Subjective and System Evaluation

We prepared 20 far-end-microphone signal pairs $\{x_j, y_j\}$ (x_j is the jth farend signal and y_j is the jth microphone signal) extracted during the video conference. The length of each signal is 10 s. Table 5 shows the MOS of each method about these 20 signal pairs. As can be seen from the table, the MOS of the proposed method is the highest. Table 6 shows the results of comparing AECMOS, RTF, network size and final score M of the eight methods. The final score M of the proposed method is the highest, and compared to the same final score, the AECMOS value is lower and the network size is larger, but the RTF is smaller.

No	method	ERLE	SAR	SI-SNR (C/N-F/N-N/N-FN)	AECMOS	RTF	network size[106]	*M*
1	Baseline [43]	17.74	14.70	12.8/8.1/1.9/1.6	3.65	0.283	5.08	0.75
2	F-T-LSTM[37]	23.98	17.56	13.5/10.4/6.2/5.8	3.95	0.410	12.8	0.79
3	Best 2022 Method A	24.10	17.80	13.8/10.5/6.3/6.0	4.00	0.60	1.5	0.85
4	Best 2022 Method B	23.85	17.65	13.6/10.2/6.2/5.8	3.97	0.10	17.4	0.84
5	Best 2022 Method C	23.45	17.40	13.5/10.2/6.1/5.7	3.92	0.20	4.8	0.82
6	Best 2022 Method D	22.80	16.95	13.3/10.0/5.9/5.5	3.85	0.30	55.5	0.80
7	Best 2022 Method E	21.50	16.20	12.9/9.7/5.5/5.1	3.72	0.02	4.3	0.78
8	Proposed MIDCCRN	24.62	17.85	13.9/10.5/6.4/6.0	3.97	0.302	4.28	0.85

Table 6: AECMOS, RTF, network size, final score M evaluation

Ablation Studies

To validate the efficiency, claim in Section 2.2.2, we compare the original four-LSTM CLSTM with our complex convolution implementation. The comparative results are summarized in Table 7.

The proposed implementation achieves equivalent performance (actually +0.04 dB) with 50% fewer activation evaluations and 8% faster training, confirming actual computational savings rather than mere operation merging.

We compare two architectural paradigms: (1) Single-Network: MIDCCRN (2) with 3× parameters (12.84M) for direct joint AEC+BNS; (2) Three-Stage (Proposed): Three cascaded MIDCCRNs (4.28M total). The performance metrics for both configurations are detailed in Table 8.

The cascaded design significantly outperforms the single-network approach (+1.44 dB SI-SNR, +6.4 dB ERLE), validating the explicit task decomposition. The single-network struggles to simultaneously model echo paths and noise characteristics, resulting in suboptimal echo suppression.

We compare temporal concatenation (proposed) versus channel-wise concatenation for MIDCCRN (1). Table 9 presents the quantitative comparison of these input concatenation strategies, demonstrating that temporal concatenation improves performance and convergence speed. Temporal concatenation improves performance and convergence speed, supporting our hypothesis that preserving frequency resolution aids cross-signal alignment learning.

Implementation	SI-SNR (dB)	Parameters	Activation Calls	Training Time (min/epoch)
Original	13.85	4.35M	12*N*	28.5
Proposed (Ours)	13.89	4.28M	6*N*	26.2

Table 7: CLSTM Implementation Comparison

Architecture	SI-SNR (dB)	ERLE (dB)	SAR (dB)	MOS
Single-Network	12.45	18.2	15.3	3.78
Three-Stage (Proposed)	13.89	24.6	17.9	4.15

Table 8: Architecture Comparison

Strategy	SI-SNR (dB)	Parameters	Convergence Epochs
Channel-wise	13.42	4.31M	38
Temporal (Proposed)	13.89	4.28M	32

Table 9: Comparison of Input Concatenation Strategies for MIDCCRN (1)

Discussion

The experimental results validate three key design decisions: First, MIDCCRN's temporal concatenation enables effective fusion of heterogeneous signals (microphone, far-end, intermediate estimates) within a unified complex domain, outperforming single-input alternatives. Second, the 50% reduction in activation calls is validated as genuine computational savings without performance degradation, crucial for edge deployment.

Third, Explicit decomposition of echo estimation, cancellation, and denoising outperforms end-to-end alternatives, supporting theoretical arguments from.

Limitations include the 125 ms algorithmic latency, which while acceptable for most conferencing scenarios, may challenge ultra-low-latency applications. Future work will explore causal attention mechanisms to reduce this further.

Conclusion

This paper presented MIDCCRN, a multi-input extension of DCCRN enabling complex-valued processing of multiple acoustic signals for joint AEC and BNS. Key innovations include temporal concatenation for multi-stream fusion, computationally efficient CLSTM implementation reducing activation calls by 50%, and a three-stage cascaded framework with multi-task training. Comprehensive experiments on the ICASSP 2022 AEC Challenge dataset demonstrate state-of-the-art performance (Final Score M=0.85) with real-time capability (RTF = 0.302) and modest model size (4.28M parameters), suitable for deployment in online communication systems.

Future directions include: (1) integration of state-space models (Mamba) for enhanced long-range temporal modeling with reduced complexity; (2) extension to multi-channel scenarios with spatial filtering; (3) joint optimization with speech recognition objectives; and (4) neural architecture search to automate stage-specific network configuration [35-41].

References

1. Bishop, A. N., Fidan, B., Doğançay, K., Anderson, B. D., & Pathirana, P. N. (2008). Exploiting geometry for improved hybrid AOA/TDOA-based localization. Signal Processing, 88(7), 1775-1791.
2. Chan, F. K., So, H. C., Ma, W. K., & Lui, K. W. (2013). A flexible semi-definite programming approach for source localization problems. Digital Signal Processing, 23(2), 601-609.
3. Cutler, R., Saabas, A., Pärnamaa, T., Loide, M., Sootla, S., Purin, M., ... & Srinivasan, S. (2021, September). INTERSPEECH 2021 Acoustic Echo Cancellation Challenge. In Interspeech (pp. 4748-4752).
4. Ho, K. C., & Le, T. K. (2023). Integrating AOA with TDOA for joint source and sensor localization. IEEE Transactions on Signal Processing, 71, 2087-2102.
5. Jia, T., Wang, H., Shen, X., Jiang, Z., & He, K. (2018). Target localization based on structured total least squares with hybrid TDOA-AOA measurements. Signal Processing, 143, 211-221.
6. Kang, X., Wang, D., Shao, Y., Ma, M., & Zhang, T. (2023). An efficient hybrid multi-station TDOA and single-station AOA localization method. IEEE Transactions on Wireless Communications, 22(8), 5657-5670.
7. Karanam, C. R., Korany, B., & Mostofi, Y. (2018, April). Magnitude-based angle-of-arrival estimation, localization, and target tracking. In 2018 17th ACM/IEEE International Conference on Information Processing in Sensor Networks (IPSN) (pp. 254-265). IEEE.
8. Kong, C., Wei, G., Li, S., Ma, P., & Li, C. (2025, June). MixSENet: A Lightweight Model for Speech Enhancement with Multi-Scale Features and Contextual Modeling. In Proceedings of the 2025 International Conference on Multimedia Retrieval (pp. 625-633).
9. Lin, L., So, H. C., Chan, F. K., Chan, Y. T., & Ho, K. C. (2013). A new constrained weighted least squares algorithm for TDOA-based localization. Signal Processing, 93(11), 2872-2878.
10. Luo, J. A., Zhang, X. P., Wang, Z., & Lai, X. P. (2017). On the accuracy of passive source localization using acoustic sensor array networks. IEEE Sensors Journal, 17(6), 1795-1809.
11. Mamun, N., & Hansen, J. H. (2024). Speech enhancement for cochlear implant recipients using deep complex convolution transformer with frequency transformation. IEEE/ACM transactions on audio, speech, and language processing, 32, 2616-2629.
12. Naseri, M., Shahid, A., & De Poorter, E. (2024). Adapting UWB AoA estimation towards unseen environments using transfer learning and data augmentation. Internet of Things, 27, 101298.
13. Natarajan, S., Al-Haddad, S. A. R., Ahmad, F. A., Kamil, R., Hassan, M. K., Azrad, S., ... & Dauitbayeva, A. (2025). Deep neural networks for speech enhancement and speech recognition: A systematic review. Ain Shams Engineering Journal, 16(7), 103405.
14. Pang, D., Wang, G., & Ho, K. C. (2024). Hybrid AOA-

TDOA localization of a moving source by single receiver. IEEE Transactions on Communications, 73(6), 4088-4104.
15. Panwar, K., Fatima, G., & Babu, P. (2022). Optimal sensor placement for hybrid source localization using fused TOA–RSS–AOA measurements. IEEE Transactions on Aerospace and Electronic Systems, 59(2), 1643-1657.
16. Qu, X., Xie, L., & Tan, W. (2017). Iterative constrained weighted least squares source localization using TDOA and FDOA measurements. IEEE transactions on signal processing, 65(15), 3990-4003.
17. Wang, G., Li, Y., & Ansari, N. (2012). A semidefinite relaxation method for source localization using TDOA and FDOA measurements. IEEE Transactions on Vehicular Technology, 62(2), 853-862.
18. Wang, Y., & Ho, K. C. (2017). Unified near-field and far-field localization for AOA and hybrid AOA-TDOA positionings. IEEE Transactions on Wireless Communications, 17(2), 1242-1254.
19. Wu, H., Chen, S., Zhang, Y., Zhang, H., & Ni, J. (2015). Robust structured total least squares algorithm for passive location. Journal of Systems Engineering and Electronics, 26(5), 946-953.
20. Wu, X., Zhao, L., & Zhu, X. (2023). Efficient semidefinite solutions for TDOA-based source localization under unknown PS. Pervasive and Mobile Computing, 91, 101783.
21. Sheng, X., & Hu, Y. H. (2005). Maximum likelihood multiple-source localization using acoustic energy measurements with wireless sensor networks. IEEE transactions on signal processing, 53(1), 44-53.
22. Xiong, W., Schindelhauer, C., So, H. C., Bordoy, J., Gabbrielli, A., & Liang, J. (2021). TDOA-based localization with NLOS mitigation via robust model transformation and neurodynamic optimization. Signal Processing, 178, 107774.
23. Xu, S. (2020). Optimal sensor placement for target localization using hybrid RSS, AOA and TOA measurements. IEEE Communications Letters, 24(9), 1966-1970.
24. Xu, Z., Li, H., Yang, K., & Li, P. (2023). A robust constrained total least squares algorithm for three-dimensional target localization with hybrid TDOA–AOA measurements. Circuits, Systems, and Signal Processing, 42(6), 3412-3436.
25. Yin, J., Wan, Q., Yang, S., & Ho, K. C. (2015). A simple and accurate TDOA-AOA localization method using two stations. IEEE Signal Processing Letters, 23(1), 144-148.
26. Yu, M., Xu, Y., Zhang, C., Zhang, S. X., & Yu, D. (2023, December). Neuralecho: Hybrid of full-band and sub-band recurrent neural network for acoustic echo cancellation and speech enhancement. In 2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU) (pp. 1-8). IEEE.
27. Zhang, Y., Zou, H., & Zhu, J. (2025). Monaural Speech Enhancement Based on Multi-Resolution Feature Analysis. IEICE Transactions on Information and Systems, 2024EDL8099.
28. Cutler, R., Saabas, A., Pärnamaa, T., Purin, M., Indenbom, E., Ristea, N. C., ... & Aichner, R. (2024). ICASSP 2023 acoustic echo cancellation challenge. IEEE Open Journal of Signal Processing, 5, 675-685.
29. Zhao, F., Zhang, C., He, S., Liu, J., & Zhang, X. (2024). Deep Echo Path Modeling for Acoustic Echo Cancellation. In Interspeech.
30. Zhang, S., Kong, Y., Lv, S., Hu, Y., & Xie, L. (2021). FT-LSTM based complex network for joint acoustic echo cancellation and speech enhancement. arXiv preprint arXiv:2106.07577.
31. Sun, Z., Li, A., Chen, R., Zhang, H., Yu, M., Zhou, Y., & Yu, D. (2024, December). SMRU: Split-and-merge recurrent-based UNet for acoustic echo cancellation and noise suppression. In 2024 IEEE Spoken Language Technology Workshop (SLT) (pp. 317-324). IEEE.
32. Wang, Y., Liu, Y., Liu, J., Niu, K., & He, Z. (2025). CAGCRN: Real-Time Speech Enhancement with a Lightweight Model for Joint Acoustic Echo Cancellation and Noise Suppression. In Proc. Interspeech 2025 (pp. 768-772).
33. Braun, S., & Valero, M. L. (2022, September). Task splitting for DNN-based acoustic echo and noise removal. In 2022 International Workshop on Acoustic Signal Enhancement (IWAENC) (pp. 1-5). IEEE.
34. Wang, Z. Q., Wichern, G., & Roux, J. L. (2021). Leveraging low-distortion target estimates for improved speech enhancement. arXiv preprint arXiv:2110.00570.
35. Yu, H., Huang, G., Gao, J., & Yan, B. (2012). Practical constrained least-square algorithm for moving source location using TDOA and FDOA measurements. Journal of Systems Engineering and Electronics, 23(4), 488-494.
36. Zhang, G., Yu, L., Wang, C., & Wei, J. (2022, May). Multi-scale temporal frequency convolutional network with axial attention for speech enhancement. In ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 9122-9126). IEEE.
37. Zheng, Y., Sheng, M., Liu, J., & Li, J. (2018). Exploiting AoA estimation accuracy for indoor localization: A weighted AoA-based approach. IEEE Wireless Communications Letters, 8(1), 65-68.
38. Zheng, Z., Zhang, H., Wang, W. Q., & So, H. C. (2019). Source localization using TDOA and FDOA measurements based on semidefinite programming and reformulation linearization. Journal of the Franklin Institute, 356(18), 11817-11838.
39. Zhu, X., Yang, J., Yang, Y., Tu, W., & Wang, Z. (2025). Lightweight real-time speech enhancement: State-space models and multi-spectral scanning techniques. Neural Networks, 191, 107805.
40. Zou, Y., Liu, H., & Wan, Q. (2017). An iterative method for moving target localization using TDOA and FDOA measurements. IEEE Access, 6, 2746-2754.
41. Zou, Y., Wu, W., & Zhang, Z. (2023). Source localization based on hybrid AOA, TDOA, and RSS measurements. IEEE Sensors Journal, 23(14), 16293-16302.

Citation: Kong, H. U., Pak, K. S., Om, C., Kim, K., UiRi, C. et. al., (2026). Multi-Input Deep Complex Convolution Recurrent Network for Joint Acoustic Echo Cancellation and Background Noise Suppression in Real-Time Speech Communication. *Adv. Intell. Robot. AI Syst. 1*(1), 01-10.