

International Journal of Reasoning-based Intelligent Systems

ISSN online: 1755-0564 - ISSN print: 1755-0556

<https://www.inderscience.com/ijris>

A deep learning-based model for noise reduction and audio quality optimisation in music transmission signals

Yang Song

DOI: [10.1504/IJRIIS.2026.10077140](https://doi.org/10.1504/IJRIIS.2026.10077140)

Article History:

Received:	30 December 2025
Last revised:	29 January 2026
Accepted:	29 January 2026
Published online:	07 April 2026

A deep learning-based model for noise reduction and audio quality optimisation in music transmission signals

Yang Song

Department of Art and Physical Education,
Huanghe University of Science and Technology,
Zhengzhou, 450000, China
Email: iamsyang@126.com

Abstract: To address noise contamination, spectral compression, and reconstruction distortion encountered in wireless transmission, remote collaboration, and embedded audio devices, this paper proposes a deep learning-based framework for audio denoising and audio quality enhancement of music-oriented transmission signals. A dual-path convolutional network incorporating frequency-domain attention and perceptually guided composite loss is then designed to model long-term noise and transient musical details simultaneously. Distinct from speech-oriented models like DCCRN, SEGAN and U-Net, the proposed method fully leverages music-specific spectral dynamics and phase consistency for high-fidelity restoration under complex distortions. Evaluated using real and simulated noise scenarios with metrics including PESQ, STOI, SI-SNR and MOS-LQO, the model achieves a mean PESQ of 3.21 and an average SI-SNR improvement of 16.4 dB, outperforming baselines. Ablation and spectral visualisation validate the key modules. With strong adaptability, the framework is applicable to real-time remote music communication, collaborative systems and high-fidelity acoustic acquisition.

Keywords: dual-path convolution; frequency-domain attention; perceptual loss optimisation; musical signal restoration; STFT-based feature modelling.

Reference to this paper should be made as follows: Song, Y. (2026) 'A deep learning-based model for noise reduction and audio quality optimisation in music transmission signals', *Int. J. Reasoning-based Intelligent Systems*, Vol. 18, No. 11, pp.29–37.

Biographical notes: Yang Song received his Bachelor of Arts and Master of Arts degrees from the Department of Chinese Instrumental Music, Wuhan Conservatory of Music in 2012 and 2016 respectively. Since September 2012, he has been working at the Music Science and Education Center, Huanghe Science and Technology College. His current research interests include instrumental performance, music education, digital audio production and AI digital audio applications.

1 Introduction

With the rapid advancement of smart audio communication, remote performance systems, and online music streaming platforms, music transmission signals are increasingly affected by background noise, bandwidth compression, multipath propagation, and codec distortion during acquisition and delivery. These factors degrade spectral integrity, distort phase structures, and consequently reduce perceived audio quality and reconstruction fidelity. Conventional filter-based audio denoising techniques rely heavily on prior noise assumptions and linear processing strategies, which limits their effectiveness in non-stationary and complex transmission environments.

Recent studies have introduced deep neural networks to enhance robustness against diverse acoustic interference. Lin et al. proposed a U-Net-based strategy to suppress wind noise in hearing devices, achieving spectral-level enhancement (Lin et al., 2024). Nogales et al. (2024) investigated the influence of multiple noise types on speech

transmission and developed a deep learning-based noise reduction framework. Lu et al. (2025) adopted a multi-model fusion approach for mixed speech-music reconstruction; however, the method suffered from over smoothing and loss of structural details. Liu et al. (2025) improved urban noise recognition using YAMNet-based transfer learning, while Han et al. (2024) combined multi-scale convolution and attention mechanisms to capture key frequency-band information in music-related tasks. Although these approaches demonstrate promising performance, most existing models are primarily optimised for speech enhancement or classification, and they do not jointly consider the time-frequency structure of music signals, perceptual audio quality, and phase-consistent reconstruction.

Music signals exhibit richer harmonic content, wider bandwidth dynamics, and stronger temporal-spectral coupling than speech signals, which makes direct adaptation of speech-oriented denoising models such as DCCRN,

SEGAN, or U-Net insufficient for high-fidelity music transmission. In particular, single-path architectures struggle to simultaneously model stationary background noise and transient musical components, while magnitude-only optimisation often neglects perceptual consistency and phase coherence.

To address these limitations, this paper proposes a deep learning-based framework that integrates dual-path convolutional modelling, frequency-domain attention, and perceptually guided composite loss optimisation for music transmission scenarios. The dual-path design separately captures long-term contextual noise and short-term transient structures, the attention mechanism emphasises semantically important frequency bands, and the perceptual loss jointly constrains spectral, phase, and waveform fidelity. This combination enables accurate noise suppression and audio quality enhancement while preserving musical details and subjective listening quality.

2 Application principles of deep learning in audio signal processing

Deep learning has emerged as a fundamental approach in modern audio signal processing, enabling the extraction of high-dimensional representations and robust signal reconstruction under complex acoustic conditions. In contrast to traditional filtering techniques – which typically rely on static noise models and prior spectral assumptions – deep neural networks are capable of learning the joint time-frequency characteristics of audio signals, including non-stationary and nonlinear noise distributions (Peng et al., 2023).

A common strategy involves transforming time-domain audio signals into complex-valued spectrograms using the short-time Fourier transform (STFT), where the resulting spectral matrix encapsulates both magnitude and phase components across frequency bins and time frames. These representations allow the network to exploit spectral energy patterns and phase dynamics simultaneously during training.

To optimise the perceptual reconstruction quality, many architectures employ composite loss functions that combine magnitude error, phase coherence, and signal-to-noise ratio (SNR)-based objectives. For instance, the training loss function may take the form shown in equation (1), where the model minimises discrepancies between the estimated and clean spectra in both magnitude and phase domains:

$$L_{total} = \lambda_1 \times \|\hat{X} - X\|^2 + \lambda_2 \times \|\angle \hat{X} - \angle X\|^2 + \lambda_3 \times \text{SDR}(x(t), \hat{x}(t)) \quad (1)$$

where $|X|$ and $\angle X$ denote the magnitude and phase of the spectrum, respectively; $\hat{X}$ is the estimated spectrum output by the network; SDR is the SNR optimisation term; and $\lambda_1, \lambda_2, \lambda_3$ is the weighting coefficient. This composite formulation ensures that the model not only improves objective metrics such as PESQ, STOI, and SI-SNR, but also enhances subjective perceptual quality by maintaining

phase fidelity and temporal consistency (Lemerrier et al., 2025).

These capabilities are particularly valuable for real-time music transmission scenarios, which often involve high dynamic range, overlapping harmonics, and low-frequency masking artifacts. Deep learning models, when properly trained, can effectively suppress interference while preserving musical semantics, outperforming conventional approaches in both intelligibility and audio quality.

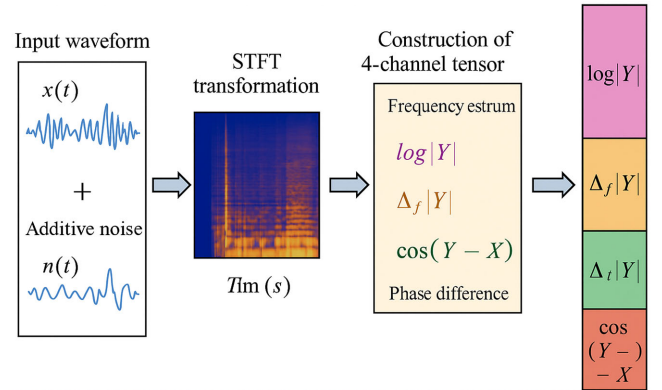
3 Modelling music transmission signal characteristics and problem analysis

During the transmission of musical content over modern digital systems, signals are subjected to a range of degradations across multiple stages, including wireless networks, codec compression, and playback hardware. These processes introduce various types of noise and nonlinear distortions, such as Gaussian background noise, low-bitrate compression artifacts, reverberation tails, packet loss, and phase jitter (Ansari et al., 2023). Accurate modelling of such degradation effects is essential for high-fidelity audio denoising and perceptual quality enhancement.

The input waveform $x(t)$ undergoes STFT transformation into the complex spectrum $X(f, t) = A(f, t) \times e^{j\phi(f, t)}$, where $A(f, t)$ represents the amplitude spectrum and $\phi(f, t)$ denotes the phase spectrum. After being affected by transmission noise $n(t)$, the received signal is equation (2):

$$y(t) = x(t) + n(t) \Rightarrow Y(f, t) = X(f, t) + N(f, t) \quad (2)$$

Figure 1 Schematic diagram of music transmission signal feature modelling structure (see online version for colours)



Considering the significant nonlinear disturbance of spectral information, a multidimensional modelling tensor must be constructed, incorporating spectral energy $\log |Y(f, t)|$, frequency band context $\Delta_f Y(f, t)$, time frame change rate $\Delta_t Y(f, t)$, and phase difference component $\angle Y(f, t) - \angle X(f, t)$ (Steinmetz et al., 2025). To characterise its statistical distribution properties under different noise types, this paper constructs a multi-channel time-frequency tensor based on the original spectral features, as shown in equation (3).

$$\mathbf{F} = [\log |Y|, \Delta_f |Y|, \Delta_t |Y|, \cos(\angle Y - \angle X)] \quad (3)$$

As shown in Figure 1, this four-channel tensor serves as the unified feature input to the deep learning model. It enables the network to capture both frequency-localised and time-varying structures across harmonic and transient components, which is especially crucial for preserving musical details during denoising and audio quality restoration. This representation forms the foundation for subsequent dual-path modelling, attention-based feature enhancement, and spectrum-phase reconstruction.

4 Design of deep learning noise reduction and audio quality optimisation models

4.1 Overall model framework

To address audio denoising and audio quality enhancement in music transmission scenarios, this study develops a deep learning framework based on a cascaded multi-module architecture that jointly integrates time-frequency representation, multi-scale context modelling, and waveform reconstruction. As illustrated in Figure 2, the system comprises an input representation layer, a feature encoder, a dual-path denoising backbone, an audio quality enhancement decoder, and a temporal reconstruction stage, forming an end-to-end spectrum-to-waveform processing pipeline.

The input waveform is first transformed using STFT to construct a four-channel spectral tensor, including the log-magnitude spectrum, frequency-difference feature, temporal-difference feature, and phase cosine difference. This representation preserves both amplitude and phase information while enhancing sensitivity to transient and reverberant distortions (Mensah et al., 2025). A residual convolutional encoder is then employed to extract stable structural components and noise-related variations from the time-frequency domain.

To effectively model heterogeneous interference in music signals, a dual-path fusion strategy is adopted. Path A utilises stacked dilated convolutions to capture long-term spectral dependencies and stationary background noise, whereas Path B integrates recurrent units and channel attention to characterise short-term transients and fine-grained harmonic dynamics. The complementary outputs of both paths are fused to produce a unified denoised representation (Wang et al., 2024). Compared with single-path encoder-decoder structures, this design enables simultaneous modelling of steady and non-stationary disturbances that commonly occur in music transmission.

The decoder incorporates multi-scale residual upsampling and channel attention blocks to recover high-frequency details and harmonic structures. The enhanced spectrum is subsequently reconstructed into the time-domain waveform via inverse STFT. During training, a composite loss function jointly constrains spectral magnitude, phase consistency, and waveform fidelity, as defined in equation (4), where the estimated spectrum, clean

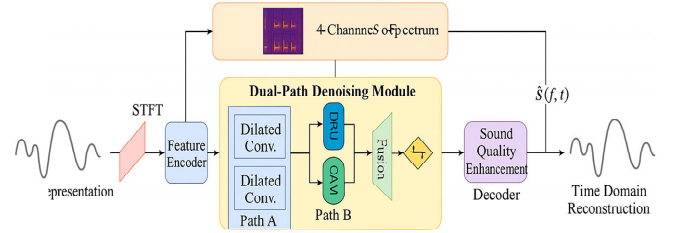
reference spectrum, reconstructed waveform, and weighting coefficients are optimised simultaneously.

$$\begin{aligned} L_{total} = & \lambda_1 \times \|\hat{S}(f, t) - |S(f, t)|\|^2 \\ & + \lambda_2 \times \|\angle \hat{S}(f, t) - \angle S(f, t)\|_1 \\ & + \lambda_3 \times (1 - \text{SI-SNR}(x(t), \hat{x}(t))) \end{aligned} \quad (4)$$

Among these, $\hat{S}(f, t)$ represents the model-estimated spectrum, $S(f, t)$ denotes the reference clean spectrum, $x(t)$, $\hat{x}(t)$ corresponds to the time-domain waveform signal, and $\lambda_1, \lambda_2, \lambda_3$ serves as the loss weighting factor. This loss function will be utilised for performance evaluation in subsequent experiments.

In addition, the model inference path supports a sliding window streaming input and a low latency reconstruction mechanism, which is suitable for low bandwidth, high interference music transmission. The proposed model is able to run in real time on the edge by parameter sharing and channel pruning, and it is able to achieve high fidelity and deployment efficiency in order to satisfy the demand of an audio service system (Kumsawat and Vanit-Anunchai, 2025).

Figure 2 Model architecture diagram (see online version for colours)



4.2 Signal feature extraction module

The feature extraction module serves as the front-end component of the proposed framework, transforming the four-channel time-frequency tensor into a compact and discriminative high-dimensional representation. The quality of this stage directly affects the effectiveness of subsequent audio denoising and audio quality enhancement modules. To achieve robust spectral modelling, the module integrates convolutional encoding, residual learning, and multi-scale channel attention to capture both local and global frequency dependencies.

As illustrated in Figure 3, the input spectral tensor first passes through a 3×3 convolution followed by batch normalisation and PReLU activation to generate an initial feature embedding (Song et al., 2023). Residual convolution blocks (RCBs) are subsequently stacked to deepen the network while preserving stable gradient propagation. Each RCB consists of two convolution-batch normalisation-PReLU sequences with shared channel dimensions and an identity shortcut connection. This residual design facilitates efficient feature reuse and enhances extraction stability without increasing

computational complexity. The corresponding feature mapping process is formulated in equation (5):

$$\mathbf{Z}_l = \sigma \left(\text{BN} \left(W_2 * \sigma \left(\text{BN} \left(W_1 * \mathbf{Z}_{l-1} \right) \right) \right) + \mathbf{Z}_{l-1} \right) \quad (5)$$

where $\mathbf{Z}_l$ denotes the output feature tensor of the RCB at layer l , $*$ represents the convolution operation, W_1, W_2 is the convolution kernel, $\sigma(\cdot)$ is the PReLU activation function, and 'BN' is batch normalisation.

To enlarge the receptive field and improve contextual perception, a multi-scale strategy is introduced through parallel convolution paths (PCP). Convolution kernels of different sizes (e.g., 3×3 , 5×5 , and 7×7) operate simultaneously to capture fine-grained harmonics as well as long-range spectral correlations. The resulting features are concatenated along the channel dimension and projected into a unified embedding space, as defined in equation (6) (Gao et al., 2024).

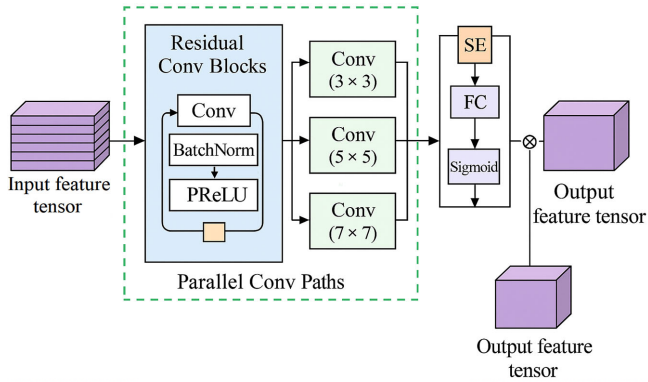
$$\mathbf{Z}_{fused} = \text{Conv}_{1 \times 1} \left(\left[\mathbf{Z}^{(3 \times 3)}; \mathbf{Z}^{(5 \times 5)}; \mathbf{Z}^{(7 \times 7)} \right] \right) \quad (6)$$

Furthermore, a squeeze-and-excitation (SE) mechanism is incorporated to adaptively recalibrate channel responses. Global average pooling first compresses the feature tensor into a channel descriptor, which is then processed by two fully connected layers with sigmoid activation to generate attention weights. These weights are applied through element-wise multiplication to emphasise frequency bands that are strongly correlated with target musical components while suppressing interference-dominated channels, as expressed in equation (7):

$$\alpha = \text{Sigmoid} \left(W_2 \times \text{ReLU} \left(W_1 \times \text{GAP} \left(\mathbf{Z}_{fused} \right) \right) \right) \quad (7)$$

The final embedded features $\mathbf{Z}_{final} = \alpha \odot \mathbf{Z}_{fused}$ are produced, where $\odot$ denotes an element-wise multiplication operation across the channel dimension, emphasising feature subspaces highly correlated with the target musical components (Richard et al., 2023).

Figure 3 Feature extraction module architecture (see online version for colours)



The whole feature extraction module achieves effective cross-scale and cross-channel modelling by convolution – residual – attention synergy. A uniform high dimensional representation is constructed to be used in the following two-path de-noising module, which provides a solid basis for the final sound quality restoration.

4.3 Multi-scale denoising module

To effectively address diverse noise patterns present in music signals, a dual-path audio denoising module with a multiscale parallel structure is constructed. This module explicitly separates the modelling of stationary and non-stationary interference, thereby improving temporal-spectral resolution and enhancing robustness against both background noise and transient distortions (Jadhav et al., 2023). It forms the core of the model's backbone and receives high-dimensional feature embeddings from the preceding feature extraction module.

Path A utilises stacked dilated convolutions to capture long-term contextual dependencies and model stationary components such as low-frequency background noise. In contrast, Path B integrates dynamic recurrent units (DRUs) and an onset-aware module (OAM) to learn short-duration variations, particularly relevant for non-stationary interference such as percussion noise and musical transients. The two branches output spectral representations $\mathbf{Z}_A$ and $\mathbf{Z}_B$, which are fused through a learnable gated weighting mechanism to form a unified denoised representation, as defined in equation (8).

$$\mathbf{Z}_{denoise} = \gamma \times \mathbf{Z}_A + (1 - \gamma) \times \mathbf{Z}_B \quad (8)$$

Here, $\gamma \in [0, 1]$ serves as a dynamically learnable gating factor to balance the contribution ratios of both paths. To support varied deployment requirements, three configuration profiles are defined: lightweight (L), standard (S), and high-fidelity (H). Each configuration differs in convolutional depth, recurrent complexity, and attention granularity, as summarised in Table 1 (Jabez et al., 2025). Configuration L is optimised for edge devices with low computational budgets; S offers a balanced trade-off for general applications; and H provides enhanced denoising fidelity through expanded context modelling and advanced attention mechanisms.

This modular design allows the model to decouple temporal and spectral denoising processes, supporting flexible adaptation to real-time constraints and fidelity requirements across different musical transmission scenarios.

Table 1 Multi-scale denoising module parameter configuration

Configuration ID	Path A convolution layers	Path B unit type	OAM channel count	Parameters (M)
L (lightweight)	4	GRU	16	0.93
S (standard)	6	DRU	32	2.85
H (high fidelity)	8	DRU + CBAM	48	5.71

Table 2 Sound quality enhancement and reconstruction module configuration

Configuration ID	Upsampling method	Phase estimation strategy	CBAM implementation	Maximum supported sampling rate (kHz)	Real-time processing capability
R1 (basic)	Linear interpolation + Convolution	None	No	16	Yes
R2 (enhanced)	Transposed convolution + Convolution fusion	Phase alignment estimator	Yes	24	Yes
R3 (high fidelity)	Multi-scale upsampling stack	UNet-based phase network	Yes	48	No (offline)

4.4 Sound quality enhancement and reconstruction module

To ensure perceptually faithful signal restoration after denoising, an audio quality enhancement and reconstruction module is introduced following the multiscale denoising stage. This module combines frequency-domain enhancement with time-domain waveform reconstruction to generate an end-to-end high-quality output. The architecture consists of a spectral enhancement sub-network, a decoder upsampling module, and a waveform synthesis mechanism.

The spectral enhancement component employs a multi-scale convolutional structure with cross-channel residual connections and frequency-band attention. These are specifically designed to recover high-frequency harmonic details and transient artifacts commonly attenuated during denoising. The enhanced spectrogram is then upsampled via a multi-stage transposed convolution decoder to match the original resolution. Subsequently, it is fused with estimated phase information and converted into the time-domain waveform using inverse STFT (Huang et al., 2023). The overall process is represented in equation (9):

$$\hat{x}(t) = \text{ISTFT}(\hat{A}(f, t) \times e^{j\hat{\phi}(f, t)}) \quad (9)$$

Among these, $\hat{A}(f, t)$ represents the enhanced amplitude spectrum, $\hat{\phi}(f, t)$ denotes the estimated phase information, and $\hat{x}(t)$ signifies the final synthesised waveform signal.

To guide perceptual consistency, a composite loss function is used during training, combining magnitude error, phase deviation, and waveform-level perceptual loss. Objective metrics such as PESQ, STOI, SI-SNR, and MOS-LQO are employed during evaluation to quantitatively assess audio quality improvements (Imon and Vaida, 2025; Szymoniak and Kuczyński, 2025).

This module supports three configurations, as summarised in Table 2, designed for deployment scenarios with varying fidelity and real-time constraints. The R1

configuration prioritises computational efficiency using simple interpolation; R2 introduces phase estimation and CBAM attention for enhanced detail; and R3 adopts a UNet-based phase network and multi-scale upsampling, targeting offline high-fidelity audio processing.

This reconstruction framework ensures structural consistency in both spectral and temporal domains, making it suitable for music-oriented applications where accurate harmonic and phase preservation are critical for perceptual realism.

5 Model training mechanism and loss function optimisation

To ensure convergence stability and optimise task-specific performance, a two-phase supervised learning strategy is adopted. In the initial stage, only the feature extraction and denoising modules are trained, with the objective of minimising spectral reconstruction error. Once convergence is observed, the audio quality enhancement and reconstruction modules are integrated, and a joint optimisation strategy is applied.

Training data are constructed by mixing clean musical signals with diverse noise types (e.g., environmental noise, reverberation, codec artifacts), with SNRs ranging from −5 dB to +15 dB. All inputs are normalised to a 16 kHz sampling rate and converted to a 1024-point STFT using a 25 ms Hamming window and 10 ms hop size. To enhance model robustness, spectro-temporal augmentation techniques – band masking, temporal masking, and amplitude perturbation – are applied.

The AdamW optimiser is used for weight updates, with an initial learning rate of 1×10^{-4} , decayed by a factor of 0.5 every 10 epochs until reaching a floor of 1×10^{-6} . Gradient clipping is applied with a maximum norm threshold of 1.0 to prevent exploding gradients. Weight initialisation is performed using He normal distribution to maintain a balanced learning dynamic across layers (Vijay et al., 2024).

The loss function consists of the amplitude spectrum loss, the phase structure loss, and the perceptual loss of the waveform restoration. Spectral domain loss is the main limitation of the model's ability to recover spectral energy, and the mean squared error is the metric. The phase structure term is used to measure the consistency between the phase and the reference signal via cosine similarity, which improves the accuracy of the restoration of complex domain. While the waveform-aware component incorporates perceptual scoring functions (for example, SDR and PMSQE) between the target and the estimated signal as guidance for improving the subjective audio quality measurement. A loss weighting mechanism is used during the training, which sets adaptive balancing coefficients for each objective. The initial coefficients are determined on the basis of the average task response gradient and are re-assessed at every epoch (Jannu and Vanambathina, 2025).

A multi-GPU parallel training scheme is used to train a batch size of 24 with 150 training cycles. The early termination mechanism is triggered when the model does not improve over a period of 10 successive epochs in the PESQ metrics, while retaining the current optimum parameters for later testing and deployment. All of the training logs, convergence curves, and loss tracking data will be used to analyse the performance of the system, and will be a key reference for future experiments.

6 Experimental setup and results analysis

6.1 Experimental setup

To evaluate the performance of the proposed model comprehensively, a combination of public datasets and synthesised samples was used. Clean speech data were sourced from the VoiceBank and VCTK corpora, while musical content was drawn from the GTZAN Genre Collection. All audio files were resampled to 16 kHz mono and normalised to 4-second durations via truncation or zero-padding to ensure temporal consistency across batches.

To simulate real-world signal degradation, three categories of distortions were introduced:

- 1 additive noise – including industrial, traffic, and instrumental sources – with randomly selected SNR values in the range $[-5 \text{ dB}, +15 \text{ dB}]$
- 2 reverberation effects generated via room impulse responses (RIRs) synthesised using the image method
- 3 compression artifacts introduced by bidirectional encoding through MP3-L2 and AMR codecs to emulate bandwidth-limited and mobile transmission scenarios.

During model training, spectro-temporal augmentation strategies – random bandwidth clipping, temporal masking, and spectro-phonetic perturbation – were employed to enhance robustness under varying acoustic conditions. Temporal masking simulates speech occlusion, while

amplitude perturbation encourages invariance to dynamic range variations.

For baseline comparison, three representative models were selected:

- 1 traditional spectral subtraction filtering (SS)
- 2 DCCRN, a time-domain convolutional recurrent network designed for real-time speech enhancement
- 3 SEGAN, a generative adversarial network for end-to-end audio restoration.

All baseline models were implemented using official codebases or publicly verified reimplementation. The dataset was split into 80% training, 10% validation, and 10% test sets. All evaluation was performed on unseen and non-overlapping samples.

Four objective and subjective metrics were used to assess performance: perceptual evaluation of speech quality (PESQ), short-time objective intelligibility (STOI), scale-invariant signal-to-noise ratio (SI-SNR), and mean opinion score – listening quality objective (MOS-LQO). These collectively measure spectral accuracy, intelligibility, noise suppression, and perceptual quality.

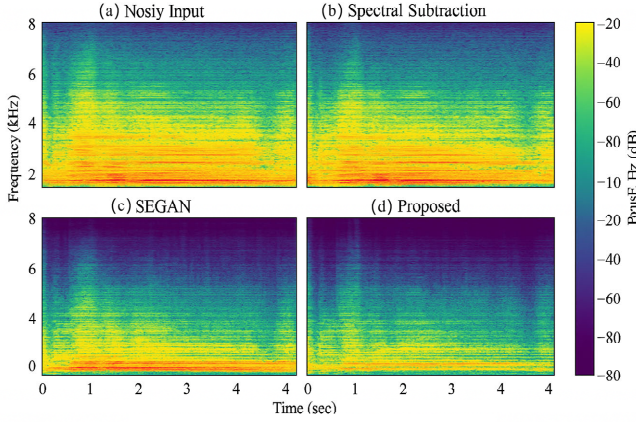
6.2 Comparison of noise reduction and sound quality enhancement effects

To assess the perceptual effectiveness of the proposed model, a comparative spectrogram analysis was conducted under consistent inference conditions. Figure 4 presents the time-frequency distributions of four enhancement strategies in a traffic noise scenario. Subfigure 4(a) shows that the noisy input contains substantial energy concentration in the 2–5 kHz range, with sustained high-energy tails masking the harmonic structure of the original signal. Subfigure 4(b) illustrates that traditional Spectral Subtraction suppresses some mid-band noise, but leaves high-frequency residuals and results in fuzzy spectral contours. In 4(c), SEGAN achieves relatively stronger denoising yet suffers from spectral over-smoothing, leading to a loss of fine-grained musical textures. In contrast, subfigure 4(d) demonstrates that the proposed model significantly enhances spectral contrast, restores harmonic boundaries, and effectively reduces both stationary and transient interference. The restored signal exhibits higher time-frequency resolution, preserving more musical details and exhibiting superior temporal coherence and noise suppression performance.

Quantitative results, summarised in Table 3, further confirm the model's superior performance across three representative distortion scenarios. In the industrial noise condition, the proposed model achieves PESQ 3.14, STOI 0.93, SI-SNR 15.7 dB, and MOS-LQO 3.91 – outperforming both DCCRN and SEGAN. Similar trends are observed under reverberant and compression distortions, with consistent gains in all four metrics. Notably, in communication-compressed conditions, the model reaches 4.10 in MOS-LQO and 16.4 dB in SI-SNR,

indicating substantial perceptual and intelligibility enhancement.

Figure 4 Pre-and post-spectral enhancement comparison across models in traffic noise scenarios (Unit: dB) (see online version for colours)



Across all configurations, the audio quality augmentation (AQA) module exhibits robust enhancement of spectral detail and intelligibility. Especially in non-stationary interference regions such as plosive consonants and percussive music segments, the model shows significant gains in perceptual and objective scores. The integrated results verify that our architecture achieves a favourable trade-off between noise suppression, spectral restoration, and perceptual fidelity.

6.3 Ablation studies and module effect analysis

To investigate the contribution of key structural components, we performed three ablation experiments by selectively removing:

- 1 the dual-path de-noising module (DP)
- 2 the frequency domain attention mechanism (ATT)
- 3 the compound loss function incorporating perceptual constraints (LF).

Each variant was trained under identical settings and evaluated in two representative scenarios: traffic noise and industrial reverberation. The full model (FULL) served as the baseline. Results in PESQ, STOI, SI-SNR, and MOS-LQO are shown in Table 4.

Among all configurations, the FULL model consistently achieves the highest performance across both noise scenarios. Removing the attention mechanism leads to the most pronounced decrease in STOI, indicating its role in enhancing frequency subbands that are semantically relevant and improving overall intelligibility. Eliminating the dual-path structure causes significant reductions in SI-SNR and MOS-LQO, underscoring the complementary roles of short-path (transient) and long-path (stationary) modelling. Single-path structures cannot fully decouple multi-scale noise distributions.

Moreover, when perceptual losses such as SDR and PMSQE are excluded, PESQ and MOS-LQO drop notably. This suggests that spectral accuracy alone is insufficient for perceptual fidelity, and that phase-aware and perceptually grounded constraints are crucial for high-quality reconstruction.

Table 3 Comparison of audio quality metrics across models under multi-noise conditions

Noise type	Model	PESQ $\uparrow$	STOI $\uparrow$	SI-SNR $\uparrow$ (dB)	MOS-LQO $\uparrow$
Industrial noise	DCCRN	2.53	0.89	13.1	3.42
	SEGAN	2.38	0.84	11.6	3.20
	Proposed model	3.14	0.93	15.7	3.91
Reverb superposition	DCCRN	2.46	0.86	12.5	3.37
	SEGAN	2.21	0.81	11.0	3.12
	Proposed model	3.02	0.91	15.2	3.85
Communication compression	DCCRN	2.59	0.90	13.5	3.48
	SEGAN	2.27	0.83	12.0	3.26
	Proposed model	3.21	0.94	16.4	4.10

Table 4 Comparison of audio quality metrics under different module configurations (traffic noise/industrial reverberation)

Model configuration	PESQ $\uparrow$	STOI $\uparrow$	SI-SNR $\uparrow$ (dB)	MOS-LQO $\uparrow$
FULL	3.21/3.02	0.94/0.91	16.4/15.2	4.10/3.85
-ATT (no attention)	2.98/2.83	0.90/0.87	14.9/13.6	3.72/3.49
-DP (single path)	2.91/2.75	0.89/0.85	13.8/12.9	3.60/3.33
-LF (No SDR/PMSQE)	2.95/2.77	0.91/0.86	14.2/13.0	3.66/3.35

In conclusion, the attention module refines local spectral focus, the dual-path structure enhances temporal noise modelling granularity, and the compound loss function ensures perceptual consistency. The synergy of these components is fundamental to the system's superior performance in both objective and subjective metrics.

7 Conclusions

Audio signal transmission is often challenged by complex acoustic interference, including non-stationary noise, high-compression artifacts, and reverberant distortions. Traditional denoising filters and single-path models generally lack robustness in such scenarios. To address these limitations, this paper proposes a novel end-to-end architecture that integrates four key modules: multi-channel spectral feature extraction, dual-path multiscale denoising, perceptual quality enhancement, and time-domain reconstruction. The proposed system leverages attention mechanisms to enhance spectral resolution and adopts perceptually-driven loss functions to improve subjective listening quality.

Extensive experiments demonstrate that the proposed model significantly outperforms baseline methods (DCCRN, SEGAN, etc.) across multiple objective metrics including PESQ, STOI, SI-SNR, and MOS. Visual analysis of spectrograms confirms the system's effectiveness in suppressing dynamic noise while preserving semantically relevant frequency components. Ablation studies further reveal the critical contributions of frequency-domain attention, dual-path design, and compound loss optimisation in maintaining both fidelity and intelligibility.

The proposed framework also exhibits strong modular scalability and adaptability to various real-time or high-fidelity applications, including speech enhancement, music restoration, and smart audio devices. However, limitations still exist. The current model incurs high computational costs and may exhibit reduced generalisability in low-SNR or highly diverse signal domains. Future work will explore lightweight modeling techniques such as parameter pruning, quantisation, and knowledge distillation to reduce inference latency. Moreover, integrating cross-modal semantic alignment modules may enhance robustness in restoring semantically degraded or incomplete acoustic signals.

Declarations

No conflict of interest exists in the submission of this manuscript.

References

Ansari, S., Alnajjar, K.A., Khater, T., Mahmoud, S. and Hussain, A. (2023) 'A robust hybrid neural network architecture for blind source separation of speech signals exploiting deep learning', *IEEE Access*, Vol. 11, pp.100414–100437.

Gao, W., Liu, Y. and Chen, D. (2024) 'A dual-stream deep learning-based acoustic denoising model to enhance underwater information perception', *Remote Sensing*, Vol. 16, No. 17, p.3325.

Han, X., Chen, F. and Ban, J. (2024) 'A GAI-based multi-scale convolution and attention mechanism model for music emotion recognition and recommendation from physiological data', *Applied Soft Computing*, October, Vol. 164, p.112034.

Huang, E.H.H., Chao, R., Tsao, Y. and Wu, C.M. (2023) 'Electrodenet – a deep-learning-based sound coding strategy for cochlear implants', *IEEE Transactions on Cognitive and Developmental Systems*, Vol. 16, No. 1, pp.346–357.

Imon, L.D. and Vaida, M.F. (2025) 'Audio denoising using U-Net architecture', *Acta Technica Napocensis*, Vol. 65, No. 1, pp.5–10.

Jabez, V.D.M., Ahilan, A., Santhaiah, C. and Thangathurai, V. (2025) 'CMANC Net: fennec fox optimized CNN-BiLSTM network for real time and complex multitude active noise cancellation', *Circuits, Systems, and Signal Processing*, pp.1–26, Early access.

Jadhav, S., Kuchibhotla, R., Agarwal, K., Habib, A. and Prasad, D.K. (2023) 'Deep learning-based denoising of acoustic images generated with point contact method', *Journal of Nondestructive Evaluation, Diagnostics and Prognostics of Engineering Systems*, Vol. 6, No. 3, p.031002.

Jannu, C. and Vanambathina, S.D. (2025) 'An overview of speech enhancement based on deep learning techniques', *International Journal of Image and Graphics*, Vol. 25, No. 1, p.2550001.

Kumsawat, N.S.P. and Vanit-Anunchai, S. (2025) 'Audio noise reduction technique using deep learning for cave rescue application', *Advances in Computer Science and Ubiquitous Computing: Proceedings of CSA 2024*, Vol. 1416, p.206.

Lemercier, J.M., Richter, J., Welker, S., Moliner, E., Välimäki, V., Gerkmann, T., et al. (2025) 'Diffusion models for audio restoration: a review [special issue on model-based and data-driven audio signal processing]', *IEEE Signal Processing Magazine*, Vol. 41, No. 6, pp.72–84.

Lin, K.Y., Lai, F.M. and Chang, H.Y. (2024) 'Wind noise reduction strategy in hearing aids through U-Net deep learning and microphone enclosure design', *IEEE Sensors Journal*, Vol. 24, No. 11, pp.18307–18316.

Liu, L.F., Xu, Q.N., Mao, S.H., Mu, J., Zhao, X., Song, W., et al. (2025) 'YAMNet-based transfer learning for compact noise classification in urban and wireless systems', *EURASIP Journal on Wireless Communications and Networking*, Vol. 2025, No. 1, p.74.

Lu, C.T., Shen, J.H., Castiglione, A., Chung, C.H. and Lu, Y.Y. (2025) 'A speech denoising demonstration system using multi-model deep-learning neural networks', *Multimedia Tools and Applications*, Vol. 84, No. 29, pp.34795–34817.

Mensah, S.Y., Zhang, T., Mahmud, N.A.I. and Geng, Y. (2025) 'Deep learning-based speech enhancement for robust sound classification in security systems', *Electronics*, Vol. 14, No. 13, p.2643.

Nogales, A., Caracuel-Cayuela, J. and García-Tejedor, Á.J. (2024) 'Analyzing the influence of diverse background noises on voice transmission: a deep learning approach to noise suppression', *Applied Sciences*, Vol. 14, No. 2, p.740.

Peng, Z., Kong, Q., Yuan, C., Li, R. and Chi, H.L. (2023) 'Development of acoustic denoising learning network for communication enhancement in construction sites', *Advanced Engineering Informatics*, Vol. 56, p.101981.

- Richard, G., Smaragdis, P., Gannot, S., Naylor, P.A., Makino, S., Kellermann, W., et al. (2023) 'Audio signal processing in the 21st century: the important outcomes of the past 25 years', *IEEE Signal Processing Magazine*, Vol. 40, No. 5, pp.12–26.
- Song, Y., Chu, Q., Liu, F., Wang, T. and Shen, T. (2023) 'Underwater acoustic signal noise reduction based on a fully convolutional encoder-decoder neural network', *Journal of Ocean University of China*, Vol. 22, No. 6, pp.1487–1496.
- Steinmetz, C.J., Uhle, C., Everardo, F., Mitcheltree, C., Mcelveen, J.K., Jot, J.M., et al. (2025) 'Audio signal processing in the artificial intelligence era: challenges and directions', *Journal of the Audio Engineering Society*, Vol. 73, pp.406–428.
- Szymoniak, S. and Kuczyński, Ł. (2025) 'Overview of modern technologies for acquiring and analysing acoustic information based on AI and IoT', *Applied Sciences*, Vol. 15, No. 12, p.6690.
- Vijay, M., Reddy, K.C., Chirudivya, G. and Harivandana, G. (2024) 'Synthesis of deep learning supported modulation signal classification model for underwater acoustic communication', *AIP Conference Proceedings*, AIP Publishing LLC, Vol. 3031, No. 1, p.020003.
- Wang, H., Zheng, H., Zhang, Z. and Wang, G. (2024) 'A deep learning-based acoustic signal analysis method for monitoring the distillation columns' potential faults', *Applied Sciences*, Vol. 14, No. 16, p.7026.