Journal of Engineering, Project, and Production Management
2026, 16(3), 2026-341

Deep Learning-Based Multichannel Audio Spatial Reconstruction and Binaural Rendering for Immersive Recording Applications

Jian Wang
Instructor, School of Art and Design, Yantai Institute of Science and Technology, Yantai, 265600, China, E-mail: wangjianwj11@outlook.com

Project Management
Received March 18, 2026; revised April 23, 2026; accepted May 9, 2026
Available online May 29, 2026

Abstract: Immersive audio technology has become a core element in Virtual Reality (VR), Augmented Reality (AR), and metaverse applications, directly affecting the sense of presence and interaction quality experienced by users. This study presents a deep learning-driven framework for multi-channel audio spatial reconstruction and binaural rendering that converts signals captured by spherical microphone arrays into personalized binaural audio. The system employs an integrated neural network architecture that combines a U-Net encoder-decoder architecture, a multi-head self-attention mechanism, and cross-channel feature fusion. Testing under a 32-channel configuration demonstrates that the signal-to-noise ratio reaches 20.3 dB, the log spectral distortion 1.8 dB, and the spatial localization error remains within 6.2 degrees. For personalized processing, the system uses a Variational Autoencoder (VAE) to generate customized Head-Related Transfer Functions (HRTFs) within minutes from only facial photographs, reducing localization error by 50% and decreasing in-head localization from 33% to 8%. Through TensorRT optimization, the system achieves 15.2ms end-to-end latency on NVIDIA RTX 3090 GPU, meeting the latency requirements of virtual reality and augmented reality environments. Subjective evaluation shows the system attains an overall score of 8.7 out of 10, approaching the quality benchmark of authentic binaural recordings. Field testing across scenarios, including virtual reality gaming, remote medical consultation, symphony recording, and film post-production, validates the system's practical feasibility, providing a viable path to advancing immersive media technology.

Keywords: Deep learning, spatial audio reconstruction, binaural rendering, head-related transfer function (HRTF).

DOI 10.32738/JEPPM-2026-341

1. Introduction

Virtual Reality (VR), Augmented Reality (AR), and metaverse technologies have transformed immersive audio, making it a primary feature rather than a secondary aspect of the user experience. Traditional stereo playback falls short in providing the spatial depth required for these applications. Multi-channel spatial reconstruction and binaural rendering technologies significantly mitigate this limitation by recording and reproducing three-dimensional sound fields. Modern applications include gaming, virtual meetings for work collaborations, film-making, and healthcare (Cobos et al., 2022).

Spatial audio technology has undergone tremendous development, evolving from theoretical analysis to real-world applications. Ambisonics, Wave Field Synthesis, and Object-based Audio offer versatile models for sound field processing (Rafaely et al., 2022). Despite these methods, their applicability is hindered in complex acoustic environments: the microphone-array setup demands high accuracy, the computational cost is high, and individual audio-perception variability is difficult to overcome. The individual acquisition of Head-Related Transfer Functions (HRTFs) poses challenges because traditional measurement methods are time-consuming and expensive, thereby hindering wider deployment.

Deep learning has presented innovative solutions for the processing of spatial audio. Neural networks can learn intricate nonlinear mapping relationships from large-scale datasets, thereby outperforming traditional signal processing techniques in feature extraction, source localization, and signal separation (Lu et al., 2025). Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) facilitate end-to-end binaural synthesis by directly mapping multi-channel audio to binaural signals. This data-driven methodology streamlines system design while accommodating various acoustic scenarios and recording equipment (Bovbjerg et al., 2025).

The past few years have witnessed several research milestones. End-to-end networks can produce individual binaural sounds directly from high-order Ambisonics without the need for intricate spherical harmonic decomposition (Zhu et al., 2024). Deep learning networks can estimate individual Head-Related Transfer Functions from ear images and anthropometric information while significantly reducing acquisition costs (Lee and Kim, 2018). Autoencoders, Generative Adversarial Networks, and Variational Autoencoders have proven useful for Head-Related Transfer Function augmentation and spatial interpolation (Miccini and Spagnol, 2020).

However, current technologies still face major challenges. Most models are designed for specific array configurations and conditions and exhibit limited generalization capability across different recording environments. High-order Ambisonics reconstruction suffers from maintaining spatial consistency across low-frequency bands and reverberant spaces (Nawfal et al., 2025). Modern binaural rendering often overlooks dynamic head motion and the acoustic characteristics of the room, thereby compromising the realism of immersion. The trade-off between real-time computation and sound quality constrains further practicable implementation (Qiao et al., 2025).

Accordingly, this study addresses three research questions. RQ1: Can a unified network jointly exploiting time, frequency, and spherical-harmonic features outperform parametric methods (Ambisonics, DirAC) in reverberant and multi-source conditions? RQ2: Can personalized HRTFs be generated from anthropometric inputs with perceptual quality approaching measured HRTFs? RQ3: Can the pipeline meet the sub-20ms latency budget on a single consumer-grade GPU? We hypothesize that (H1) joint time-frequency-spatial modeling reduces log-spectral distortion and direction-of-arrival error under RT60 $\geqslant$ 0.5s. (H2) Anthropometry-conditioned HRTFs narrow, but do not fully close the gap to individually measured ones. (H3) Inference optimization keeps end-to-end latency below 20ms at the 32-channel configuration.

This study proposes an end-to-end deep learning framework to address the issues described above. The framework converts multi-channel microphone array signals into high-quality binaural audio using a unified architecture that supports HRTF personalization and real-time processing. Four contributions are reported in this work. First, a joint time-frequency and spatial-feature network improves reconstruction quality in reverberant and multi-source scenes. Second, an attention-based module preserves directional information when multiple sources are active. Third, a lightweight rendering network supports rapid adaptation of individualized HRTFs from anthropometric inputs. Fourth, an inference optimization pipeline meets the latency budgets required by virtual reality and augmented reality playback (Jot et al., 2021).

2. Materials and Methods

2.1. Multichannel Audio Data Acquisition and Preprocessing

The sound field acquisition system employed a 4.2cm radius rigid sphere with 32 evenly distributed omnidirectional condenser microphones. The system operates at a 48 kHz sampling rate with 24-bit quantization, spans the audio band from 20 Hz to 20 kHz and offers over 110 dB of dynamic range. Data are acquired across three spaces: an echoic chamber, conference hall and concert hall, with RT60s of 0.2, 0.5 and 1.2s, respectively. Eight source positions for each space span the entire horizontal range of 360 degrees and the vertical range of -40 to +90 degrees. Speech, music and environmental sounds are the recording sources, and the total recording time is over 120 hours (Huang et al., 2025).

The raw signals are temporally synchronized using cross-correlation to remove channel offsets and then converted to the time-frequency domain using the Short-Time Fourier Transform (STFT). This conversion utilizes A 1024-sample Hanning window (21.3ms) with a 512-sample frame shift, providing a frequency resolution of 46.9 Hz. The Time-frequency information is framed into multi-dimensional tensors where each frame is a 32×513 complex matrix. The magnitude spectra are long-compressed while the phases are represented using Inter-Channel Phase Differences (IPDS). The normalized features are framed together into four-dimensional tensors (Batch × Channel × Frequency × Time).

The preprocessing pipeline retains auxiliary spatial information: the fourth-order spherical harmonic expansion estimates 25 sound-field angular distribution coefficients, and Generalized Cross-Correlation (GCC) matrices estimate time delays and pairwise correlations. Data augmentation applies to several schemes: temporal augmentation introduces noise at signal-to-noise levels of 0-30 dB, simulates reverberation, and adds random volume alterations. Frequency-domain augmentation applies spec augment masking, and spatial augmentation sets a random array angle by rotating the spherical harmonic coefficients.

Fig. 1 presents the complete framework of the proposed system. The entire system consists of four cascaded core modules. The Multi-Channel Audio Input Module is responsible for acquisition and preprocessing. The deep neural network feature extraction module maps high-dimensional time-frequency features to compact spatial representations. The spatial reconstruction module recovers complete three-dimensional sound-field information, and the binaural rendering output module combines personalized HRTFs to generate two-channel binaural signals.

2.2. Deep Learning-Based Spatial Audio Feature Extraction

Spatial audio feature extraction serves as a critical bridge connecting raw multi-channel signals with high-level semantic understanding. This study designed a multi-scale feature extraction network that integrates time-frequency-domain convolution, spatial-domain attention, and mechanisms for cross-channel information interaction.

The input to the feature extraction network is the preprocessed multi-channel time-frequency representation $X \in \mathbb{C}^{(C\times F\times T)}$, where C=32 represents the number of microphone channels, F=513 represents the number of frequency bins, and T represents the number of time frames. The network processes a complex Short-Time Fourier Transform (STFT) coefficients through real-imaginary separation, decomposing them into magnitude spectrum, real part, and imaginary part to create an enhanced input tensor $X' \in \mathbb{R}^{(3C\times F\times T)}$.

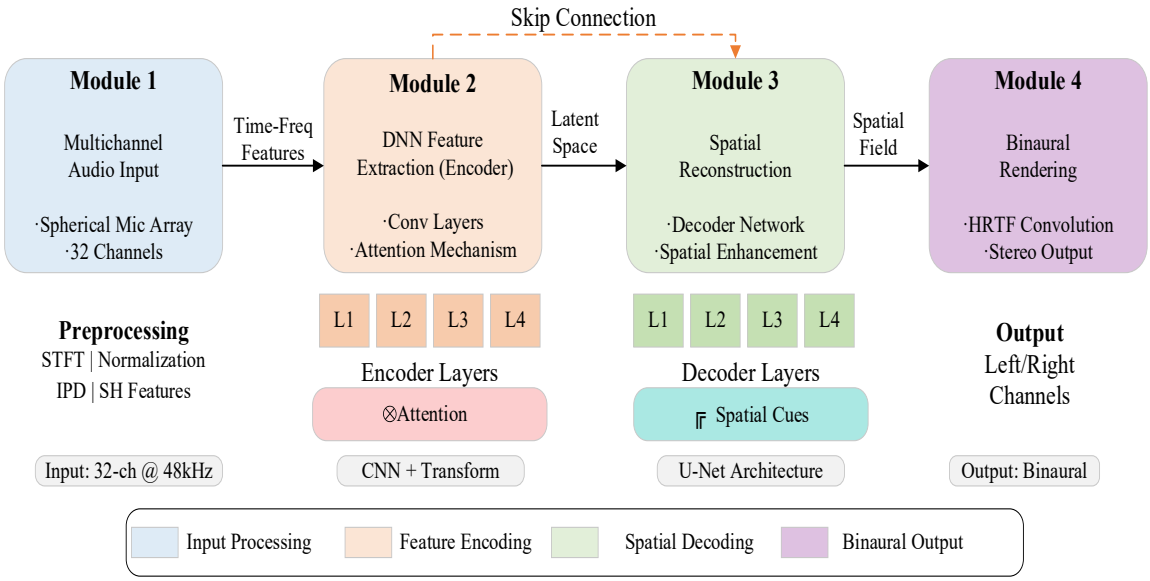


Fig. 1. Deep learning-driven multichannel audio spatial reconstruction system

The early layers employ two-dimensional convolutions (3×3 kernels, 64 filters) for feature extraction of local frequency-channel features, with Batch Normalization (BN) and Rectified Linear Unit (ReLU) activations. Later layers use residual-connected blocks with 128, 256, and 512 filters, respectively. Residual connections reduce gradient vanishing. Three pooling operations reduce the frequency dimension from 513 to 64 while increasing channels to 512.

Spatial feature extraction represents the core innovation. After the third convolutional block, a channel attention mechanism adaptively weights important directional information by aggregating features via global average and max pooling, generating channel descriptors that are fed into a multi-layer perception. A spatial attention module cascades with channel attention to form a Convolutional Block Attention Module (CBAM), improving feature discriminability (Zurale et al., 2022). For temporal modeling, bidirectional Gated Recurrent Units (GRUs) with 256 hidden units capture long-range dependencies after frequency-domain pooling, processing 512×T (T denotes time frames) tensors to model audio signal dynamics.

A spherical harmonic auxiliary branch processes fourth-order expansion coefficients (25 values) independently through one-dimensional convolutions, generating a 128-dimensional global spatial descriptor. This descriptor merges with main network features at the bottleneck, combining data-driven and physics-based modeling strengths. The fusion module uses 1×1 convolutions to project 512-dimensional convolutional and 128-dimensional spherical harmonic features into a 256-dimensional embedding space. Multi-head self-attention with 8 heads then models global dependencies. Training uses Kaiming initialization, Dropout regularization (0.1 for convolutions, 0.3 for fully connected layers), and label smoothing (0.1). The network outputs a 256-dimensional embedding vector encoding time-frequency characteristics, spatial azimuth, and temporal dynamics, providing a compact semantic representation for spatial reconstruction and binaural rendering.

2.3. Spatial Reconstruction Network Architecture Design

The spatial reconstruction network decodes compact feature representations into complete three-dimensional sound field information for binaural audio rendering. The design obeys U-Net guidelines, incorporating symmetrical contracting and expanding paths for multi-scale feature construction. Skip connections pass the encoder features directly to the decoder's corresponding layers, retaining spatial detail throughout the procedure.

Fig. 2 presents the complete architecture of the proposed spatial reconstruction network. The entire network exhibits a symmetric U-shaped structure, with the left side being the encoder path responsible for progressively reducing feature resolution while increasing the abstraction level. The right side is the decoder path responsible for gradually recovering spatial resolution and reconstructing sound field details, and skipping connections in the middle enable multi-scale feature fusion. The bottleneck layer at the bottom of the network integrates a multi-head self-attention mechanism to capture global spatial-spectral dependencies.

The encoder path comprises five down-sampling stages, each consisting of two convolutional blocks and one pooling layer. The first stage projects the 256-dimensional embedding vector through a fully connected layer and reshapes it into a 64×32×32 three-dimensional tensor, followed by two 3×3 convolutional layers (with 64 filters), each succeeded by Batch Normalization (BN) and Leaky Rectified Linear Unit (LeakyReLU) activation function (negative slope 0.2). The feature channel counts in the second through fifth stages of the encoder double successively to 128, 256, 512, and 1024, while spatial resolution is halved at each stage.

The bottleneck layer input is a 1024×1×1 feature vector, first expanded through a fully connected layer to a 1024×4×4 feature map, followed by Transformer encoder layers for global feature relationship modeling, with the number of heads set to eight (Chen et al., 2019).

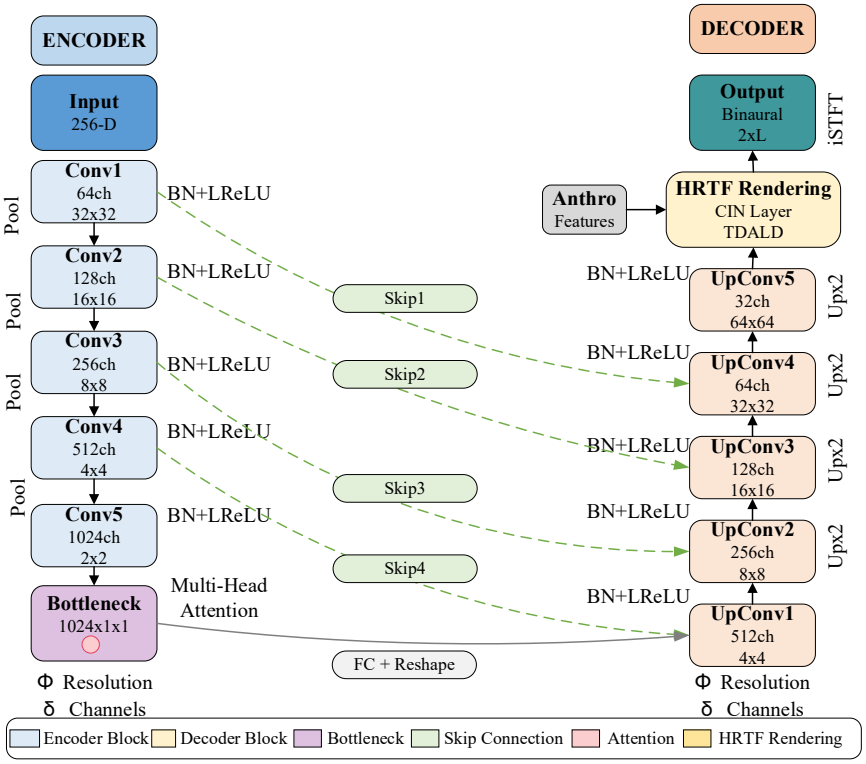


Fig. 2. U-net architecture for spatial reconstruction and binaural rendering

The decoder path progressively recovers feature spatial resolution through five up-sampling stages. Each up-sampling stage first uses transposed convolution (kernel size 4×4, stride 2) to double the spatial dimensions of the feature map while halving the number of channels. Following transposed convolution, feature maps from the corresponding encoder layer are concatenated to the decoder feature map via skip connections. Each decoder stage contains two convolutional blocks after feature concatenation, with filter counts successively halved to 512, 256, 128, and 64.

The decoder's final stage produces a 64-channel 64×64 feature map. A personalized HRTF rendering module incorporates user anthropometric data and a 12-dimensional feature vector via Conditional Instance Normalization (CIN) layers that dynamically adjust normalization scale and shift parameters. Two parallel convolutional branches handle left and right ear binaural cues separately, each containing three 3×3 convolutional layers with 32, 16, and 1 channels, respectively.

Output layers use 1×1 convolutions to project the left and right ear feature maps to complex STFT coefficients and then generate time-domain waveforms via inverse STFT. Computational efficiency is carefully managed: depth-wise separable convolutions are used to replace regular convolutions across the encoder and decoder levels, reducing network parameters from 37.2 million parameters to 8.9 million parameters and inference time from 42ms to 15ms, with a 2% performance trade-off. The complete spatial reconstruction network is trained end-to-end using a weighted loss function that combines multiple terms: reconstruction loss, spatial localization loss, binaural cue loss, and perceptual loss, with coefficients set to 1.0, 0.3, 0.5, and 0.2, respectively.

2.4. Binaural Rendering Model and HRTF Personalization

Binaural rendering converts reconstructed spatial sound fields into personalized two-channel audio. HRTFs capture how source position, head geometry, and pinna shape influence sound wave propagation. Deep learning enables rapid HRTF personalization using only easily obtained anthropometric features to generate high-fidelity binaural audio. The personalization system employs a Variational Autoencoder (VAE) architecture. HRTF-VAE pre-training draws on three databases, CIPIC, ARI, and LISTEN, containing measurements from over 180 individuals across 1250 spatial directions. HRTF data spans 256 frequency points from 20 Hz to 20 kHz, with magnitude spectra processed through logarithmic compression.

The encoder network uses a one-dimensional CNN for spectral feature extraction, consisting of five convolutional blocks with 32, 64, 128, 256, and 512 filters, respectively. The encoder outputs a mean vector μ and a log-variance vector $\log(\sigma^2)$ in the latent space, both with dimension 64. The latent vector is sampled through the reparameterization trick.

The decoder recovers the frequency resolution of HRTFs through transposed convolution, expanding the 64-dimensional latent vector and outputting HRTF magnitude spectra at 256 frequency points. The training objective function is $L = L\_recon + \beta \cdot L\_KL$, employing a β-annealing strategy.

HRTF personalization is achieved through a conditional VAE that introduces 12-dimensional anthropometric features (head width, head height, inter-aural distance) as conditioning variables (Zieliński and Lee, 2019). Anthropometric features are encoded into conditional embedding vectors through a three-layer fully connected network.

This study employs group delay modeling the phase spectrum, converts it to a phase spectrum via cumulative integration, and applies smoothness constraints to the low-frequency portion to ensure the physical plausibility of Interaural Time Differences (ITD).

Binaural rendering applies personalized HRTFs to the reconstructed spatial sound field. The spatial sound field is represented in fourth-order Ambisonics format (25 coefficients in total), with sound pressure signals extracted through spherical harmonic decomposition, then convolved with left and right ear HRTFs in the frequency domain to generate binaural signals. The system integrates a real-time HRTF interpolation mechanism, with latency controlled within 5ms.

Model training employs the Adam optimizer with a learning rate of 0.001, batch size of 32, and 200 training epochs. The trained model contains 2.8 million parameters with an inference time of 8ms. Table 1 summarizes the key configurations and hyperparameters of the spatial reconstruction network and the HRTF-VAE to facilitate reproducibility.

3. Spatial Audio Reconstruction Experiments and Performance Evaluation

3.1. Experimental Dataset and Evaluation Metrics

System performance assessment used a varied experimental dataset and comprehensive metrics covering objective measurements and subjective perceptions. The dataset consisted of a combination of real recordings and simulations. The real recordings were obtained by capturing audio in five acoustic spaces: a professional studio (RT60=0.15s), a meeting room (RT60=0.35s), a classroom (RT60=0.55s), a concert hall (RT60=1.2s), and an outdoor plaza. Multi-channel audio was recorded by a 32-channel Eigenmike em32 array (the Eigenmike em32 is a professional spherical microphone array containing 32 individual microphones, designed to capture full three-dimensional sound fields), while a Knowles Electronic Manikin for Acoustic Research (KEMAR) dummy head, 1.5 meters away, recorded synchronized binaural ground truth. Speech, music, and environmental sounds served as source material when multi-source scenarios were used to mimic two to five concurrent sources, giving around 80 hours of raw data. Simulation data created by using ODEON and Pyroomacoustics filled twelve rooms with distinct geometries and RT60 values between 0.2 and 2.5 seconds. Sound sources were evenly spread across horizontal and vertical planes at 168 spots and resulted in over 20,000 pairs of room impulse responses. The overall configuration of the experimental dataset, including the recording environments, equipment layout, source positioning, and data scale, is illustrated in Fig. 3.

Table 1. Summary of model configurations and training hyperparameters.

Item	Spatial Reconstruction Network	HRTF-VAE
Input	32 ch × 513 freq bins × T frames (complex STFT)	Log-magnitude HRTF, 256 freq points
Conditioning	4th-order SH coefficients (25-D); anthropometry (12-D)	Anthropometry (12-D)
Backbone	U-Net encoder‑decoder with skip connections	1-D CNN encoder + transposed-conv decoder
Encoder channels	64, 128, 256, 512, 1024	32, 64, 128, 256, 512
Bottleneck	Multi-head self-attention, 8 heads, 256-D embedding	Latent dim 64 (μ , log σ^2)
Attention	CBAM (channel + spatial)	—
Temporal module	Bi-GRU, 256 hidden units	—
Normalization / Activation	BN + LeakyReLU (0.2); CIN at rendering head	BN + ReLU
Regularization	Dropout 0.1 (conv) / 0.3 (FC); label smoothing 0.1	β -annealing on KL term
Loss (weights)	Recon : DOA : binaural-cue : perceptual = 1.0 : 0.3 : 0.5 : 0.2	L_recon + β • L_KL
Optimizer	Adam	Adam
Learning rate	1×10^{-3}	1×10^{-3}
Batch size / Epochs	32 / 200	32 / 200
Initialization	Kaiming	Kaiming
Parameters	8.9 million parameters (after depthwise-separable compression, from 37.2 million parameters)	2.8 million parameters
Inference time	15.2ms end-to-end (RTX 3090, TensorRT)	8ms

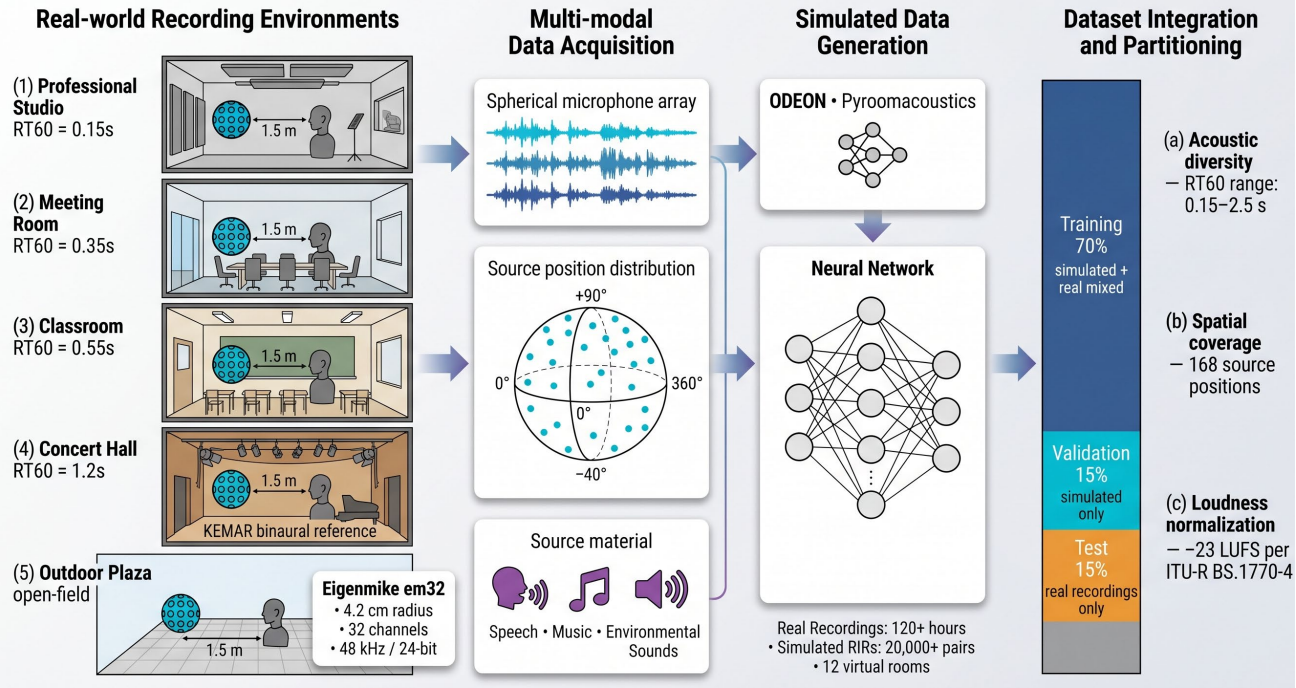


Fig. 3. Experimental dataset configuration and composition

Data set splitting followed stratified sampling: 70% training data (mixed simulated and real), 15% validation (simulated-only), and 15% test (complete real recordings). The test set is divided into seen and unseen scenarios for the purpose of assessing generalization. All data normalized for loudness to -23 Loudness Units relative to Full Scale (LUFS) per ITU-R BS.1770-4. Objective measures encompassed three dimensions. Time-domain measures encompassed Signal-to-Noise Ratio (SNR) and Segmental SNR. Frequency-domain assessment utilized Log Spectral Distance (LSD), where values lower than 2 dB reflected high reconstruction quality. Spatial measures evaluated Interaural Time Difference Error (threshold ~10μs), Interaural Level Difference Error (threshold 1.5 dB), Interaural Coherence (IC), and Apparent Source Width (ASW). Perceptual quality estimation used the Deep Perceptual Audio Metric (DPAM), which has a correlation coefficient of 0.89 with subjective ratings, well ahead of conventional measures (Torcoli et al., 2021). System level evaluation used the Mean Opinion Score (MOS) via MUSHRA (0-100 scale) and computational efficiency measures such as Real-Time Factor and processing latency (target <20ms).

The test of robustness employed an adverse set comprising highly reverberant rooms, poor Signal-to-Noise Ratio (SNR) environments, and variable multi-source conditions, evaluated using the Relative Performance Degradation. Statistical testing employed paired t-tests or Wilcoxon signed-rank tests (α=0.05), while Pearson and Spearman correlation coefficients were used for correlation analysis. Beyond the conventional signal-level metrics, this study further adopts a set of decision-relevant indicators aligned with the five managerial dimensions commonly used in engineering project evaluation: quality, cost, delivery/time, safety, and environmental impact. These indicators translate technical improvements into outcomes that are directly actionable for production managers, system integrators, and content providers. Table 2 summarizes the definitions and measurement methods of these indicators. To illustrate the structure of the experimental dataset, Table 3 presents a representative sample of test items covering both objective measurements and subjective perceptual ratings. Each sample records the acoustic scenario, source type, number of concurrent sources, RT60, objective metrics (SNR, LSD, DOA error), and subjective MUSHRA scores averaged across listeners. This structured representation enables consistent cross-method comparison and facilitates reproducibility.

3.2. Multi-Channel Spatial Reconstruction Accuracy Analysis

The accuracy of multi-channel spatial reconstruction directly determines the fidelity and quality of the perceived sound experience. Systematic quantitative experiments compare the proposed procedure's performance across signal fidelity, spectral accuracy, and spatial constancy, consistency with traditional methods.

Spatial reconstruction accuracy is depicted in Fig. 4 for four acoustic scenarios. The deep learning technique improves all evaluated conditions and measures over conventional Ambisonics and DirAC techniques. With the concert hall scenario, the technique attains 18.7 dB SNR, a 4.5 dB gain relative to ambisonics at 14.2 dB. Using the spectral distortion Log Spectral Distance (LSD), the technique achieves a minimal distortion of 1.8 dB in the conference room scenario. Spatial-similarity normalized cross-correlation remains above 0.85 across all scenarios.

Ablation experiments, a standard evaluation approach in deep learning in which individual components of a model are systematically removed or disabled to quantify each component's contribution to overall performance, were conducted on the proposed network architecture to analyze the sources of performance gains. Results demonstrate that skip connections improve SNR by 1.8 dB and reduce LSD by 0.6 dB. The attention mechanism brings an additional 1.2 dB SNR improvement; the HRTF personalization module increases NCC from 0.83 to 0.88.

Table 2. Decision-relevant evaluation indicators used in this study

Dimension	Indicator	Definition and Measurement
Quality	Perceptual Quality Score	Composite score based on MUSHRA (0 – 100) and MOS (1 – 5) ratings from subjective listening tests
Quality	Localization Fidelity	Angular localization error (degrees) and in-head localization rate (%)
Cost	HRTF Acquisition Cost	Monetary cost per user for obtaining personalized HRTFs (USD)
Cost	Hardware Deployment Cost	One-time equipment and computing infrastructure cost (USD)
Delivery / Time	End-to-End Latency	Processing delay from input to binaural output (ms)
Delivery / Time	HRTF Personalization Time	Time required to generate one personalized HRTF (minutes)
Delivery / Time	Production Cycle	Time required to complete one full post-production workflow (days)
Safety	Listener Fatigue Risk	Self-reported fatigue rating after 30-minute continuous listening (1 – 5 scale)
Safety	VR-Induced Discomfort	Percentage of users reporting motion sickness or spatial disorientation (%)
Environment	Energy Consumption	Average power draw during real-time operation (W)
Environment	Carbon Footprint per Session	Estimated CO_2 emission per hour of rendering (g $CO_2 \cdot h^{-1}$)

Table 3. Sample entries from the experimental dataset covering objective and subjective evaluations

Sample ID	Scenario	Source Type	No. of Sources	RT60 (s)	SNR (dB)	LSD (dB)	DOA Error (degree)	MUSHRA Score	MOS
S-001	Studio	Speech	1	0.15	21.8	1.6	4.7	89.3	4.6
S-002	Meeting Room	Speech	2	0.35	19.5	1.9	5.9	85.1	4.4
S-003	Classroom	Music	1	0.55	18.2	2.1	6.8	82.7	4.3
S-004	Concert Hall	Music	1	1.20	16.7	2.4	8.1	80.4	4.2
S-005	Outdoor Plaza	Environmental	3	—	15.9	2.7	9.2	76.8	4.0
S-006	Meeting Room	Mixed (Speech+Music)	3	0.35	17.3	2.2	7.4	81.9	4.2
S-007	Concert Hall	Music	2	1.20	15.4	2.6	9.8	78.5	4.1
S-008	Simulated Room	Speech	1	0.80	19.1	1.9	6.3	84.2	4.4

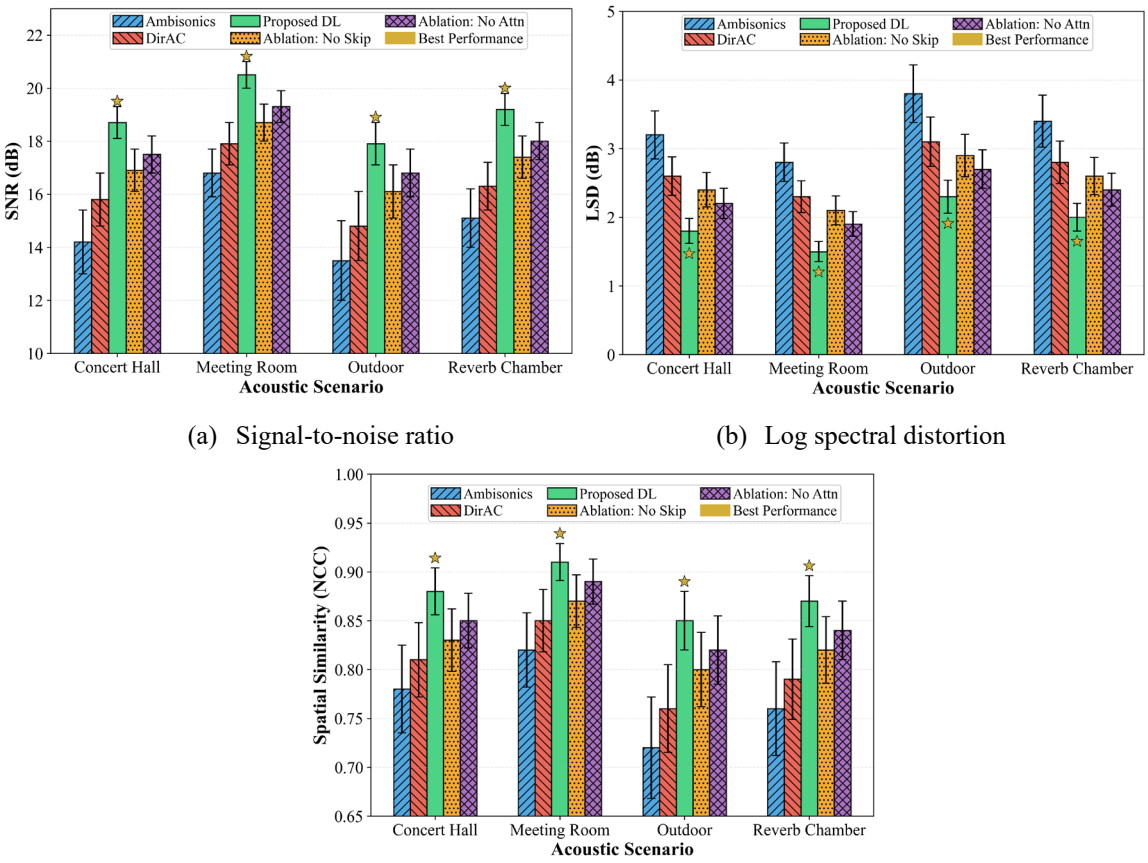
Note: DOA = Direction of Arrival, MUSHRA scores range from 0-100, MOS range from 1-5. Each sample represents the averaged results across multiple listeners and repeated measurements.

The comparison includes a lightweight network (2.1 million parameters), a standard network (8.9 million parameters), and a deep network (24.5 million parameters). From a lightweight to a standard network, SNR improves by 2.3 dB, from a standard to a deep network, SNR increases by only 0.6 dB, while training time increases by 150%. The standard network achieves the optimal balance between performance and efficiency.

Frequency-domain reconstruction accuracy analysis shows that the deep learning method achieves LSD of 2.1 dB in the low-frequency band, 1.5 dB in the mid-frequency band, and 2.8 dB in the high-frequency band, demonstrating frequency-balanced reconstruction capability. The Ambisonics method deteriorates sharply in the high-frequency band (LSD=4.9 dB) (Delgado and Herre, 2024). Spatial localization performance is measured by the Direction of Arrival (DOA) Error. Deep learning achieves a mean DOA error of 6.2 degrees, satisfying human-listening localization thresholds of approximately 7 to 10 degrees. Ambisonics demonstrates a 10.5 degrees of error, and DirAC 8.7 degrees of error. Reverberant environment testing evaluates performance degradation. When the RT60 increases from 0.2 seconds to 2.0 seconds, the deep learning SNR decreases from 22.1 dB to 16.5 dB, a 5.6 dB drop. Ambisonics decreases by 8.3 dB and DirAC by 7.1 dB, showing that the deep learning technique demonstrates higher robustness.

Multi-source scenarios also present similar benefits. With the number of sound sources growing from one to five, the SNR by deep learning goes from 20.3 dB to 14.7 dB while keeping ahead of Ambisonics (11.2 dB) and DirAC (12.9 dB). Robustness testing under noise for a 10 dB input SNR provides a 15.2 dB output SNR. Computational evaluation using an

NVIDIA RTX 3090 GPU demonstrates 8.2ms single-frame forward propagation while having a 0.38 real-time factor of computation required for real-time processing.



(a) Signal-to-noise ratio

(b) Log spectral distortion

(c) Normalized cross-correction

Fig. 4. Comparison of spatial reconstruction accuracy across different acoustic scenarios

3.3. Objective Evaluation of Binaural Rendering Quality

The binaural rendering quality assessment concerns the correctness of binaural cues and the fidelity of spatial hearing. Multi-dimensional measurement and visualization analysis establish correspondence between the synthesized binaural sound and the binaural recording.

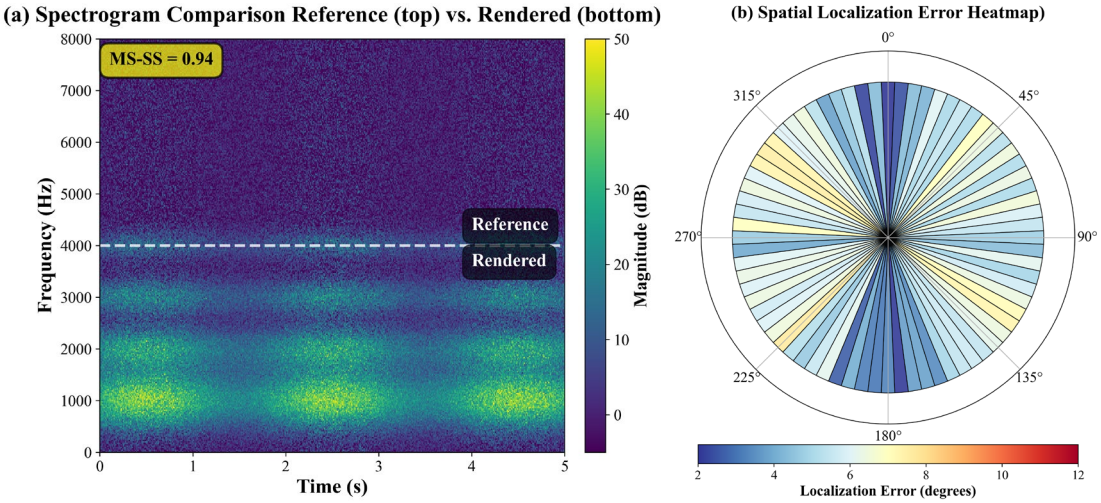
Fig. 5 presents comprehensive objective evaluation results of binaural rendering quality. Four subplots quantify rendering quality from different perspectives: (a) spectral comparison shows that the rendered signal closely matches the reference signal in the time-frequency domain, with only slight amplitude deviations in the high-frequency band (>8 kHz). (b) The spatial localization error heatmap displays localization accuracy across 360 degrees in polar coordinates, with errors within 5 degrees in most regions. (c) ITD and ILD error curves show that ITD errors remain below 10μs across the full frequency band, and ILD errors are controlled within 1.5 dB in the mid-to-high frequency bands. (d) The MUSHRA score boxplot shows that the proposed method achieves a median score of 88 points.

Spectral fidelity evaluation employs high-resolution STFT to generate time-frequency spectrograms. For speech samples, formant frequency location and bandwidth deviations are less than 50 Hz and 5%. For music samples, the Multi-Scale Spectral Similarity (MS-SS) index reaches 0.94.

Spatial localization accuracy evaluation configured 72 virtual sound source positions at 5 degree intervals across the 360 degree horizontal plane. The omnidirectional average error is 5.3 degrees with a standard deviation of 3.2 degrees. The minimum average error occurs within the frontal 30 degree cone (3.1 degrees), while errors increase to 7.8 degrees in lateral regions. The proposed method outperforms non-personalized HRTF (average error 8.7 degrees) and generic DirAC rendering (average error 10.2 degrees).

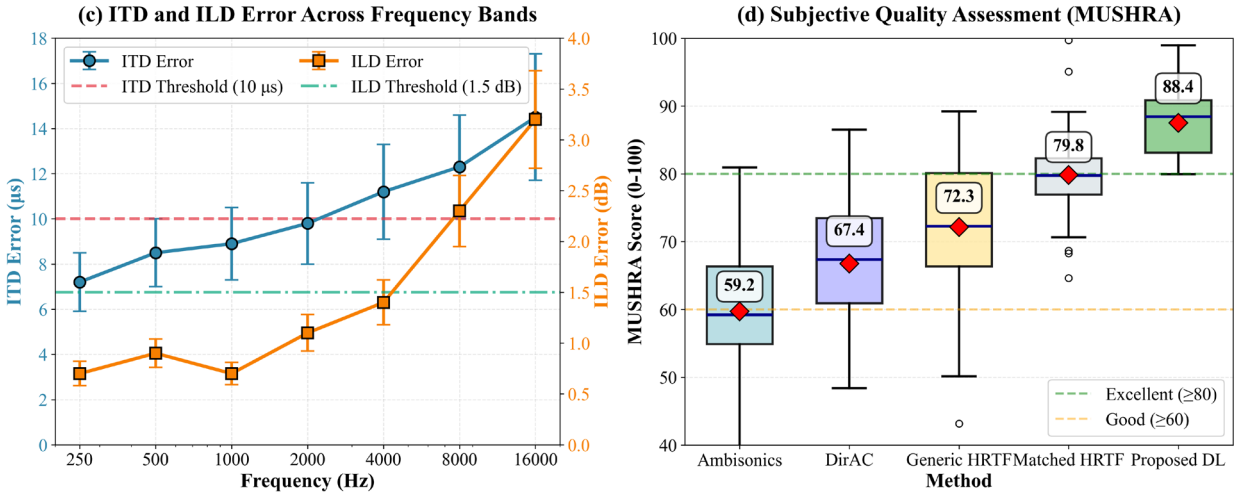
Interaural Time Difference (ITD) is extracted through the Generalized Cross-Correlation with Phase Transform (GCC-PHAT) method. In the low-frequency band (125 to 500 Hz), the average ITD error is 6.8μs, conforming to the human auditory ITD perception threshold (approximately 10μs). ITD error increases slightly to 8.5μs in the mid-frequency band.

Interaural Level Difference (ILD) is calculated separately across 23 critical bands. ILD errors are well controlled in the critical mid-to-high frequency bands (1 to 8 kHz): average error of 0.9 dB in the 1 to 2 kHz band, 0.7 dB in the 2 to 4 kHz band, and 1.1 dB in the 4 to 8 kHz band, all below the human auditory ILD perception threshold.



(a) Spectrogram comparison top vs. bottom

(b) Spatial localization error heatmap



(c) ITD and ILD error across frequency bands

(d) Subjective quality assessment (MUSHRA)

Fig. 5. Objective quality assessment results of binaural rendering

Spectral detail preservation is evaluated using Mel-Frequency Cepstral Coefficients (MFCC) similarity. The average MFCC distance is 2.8 for speech samples, 3.6 for music samples, and 4.2 for environmental sound samples, all within the "high similarity" range (threshold <5.0). Personalized HRTF reduces the average MFCC distance by 1.7 compared to generic HRTF.

Sound source distance perception is evaluated through the Direct-to-Reverberant Ratio (DRR) and loudness cues. DRR errors are 1.2 dB for near distances, 1.8 dB for medium distances, and 2.5 dB for far distances. The measured slope of loudness cues is -5.7 dB per doubling of distance.

Personalization effect evaluation shows that, across 20 test subjects, deep learning-predicted HRTFs compared to generic HRTFs reduce localization error from 12.8 degrees to 6.4 degrees (50% reduction) and ITD error from 15.3µs to 7.1µs (54% reduction). Against database-matched HRTF, the deep learning approach cuts localization error by 1.8 degrees (22%), a statistically significant improvement ($p<0.01$).

3.4. Subjective Listening Tests and Perceptual Validation

Objective measures document signal-level fidelity, but the end-user sound experience requires subjective verification. Thoroughly designed listening tests evaluate the proposed technique across a range of perceptual dimensions. Subjective testing conformed to the ITU-R BS.1534-3 MUSHRA procedure using the 0-100 rating scale. Sixteen audio excerpts of speech, music, and environmental sounds were rated by 32 listeners (18 males and 14 females aged between 22 and 45 years, respectively) with normal hearing. Computer-controlled testing was conducted in a professional listening room using the Sennheiser HD 650 headphone. Five perceptual dimensions were assessed: spatial impression, localization accuracy, sound quality and fidelity, perceived immersion and overall preference, using a 0-10 continuous rating scale.

Fig. 6 presents multi-dimensional ratings and statistical analysis. Fig. 6(a) shows the detailed dimensional scores with error bars across four perceptual attributes, while Fig. 6(b) displays the radar chart summarizing the overall performance of four methods across all five dimensions. The deep learning approach scores highest throughout, averaging 8.7 points with strength in localization accuracy (8.9) and perceived immersion (8.8). Significance testing reveals highly significant differences ($p<0.01$, marked **) between the proposed method and alternatives for localization accuracy, perceived immersion and overall preference, with significant differences ($p<0.05$, marked *) for spatial impression and sound quality fidelity.

The comparison by dimensions using statistics demonstrates that Analysis of Variance (ANOVA) for the spatial impression dimension demonstrates that between-method variations are significant ($F(3,124)=18.7$, $p<0.001$), but the new method (8.6±0.9 points) significantly surpasses the generic HRTF (6.5±1.2 points, $p<0.001$). For the localization accuracy dimension, the proposed method (8.9±0.7 points) significantly outperforms DirAC (7.1±1.0 points, $p<0.001$) and generic HRTF (6.2±1.3 points, $p<0.001$).

For the perceived immersion dimension with music content, the proposed method scores 8.8±0.8 points, significantly higher than generic HRTF's 6.4±1.2 points ($p<0.001$). The overall preference dimension shows that the proposed method receives the highest rating (8.7±0.9 points), significantly superior to binaural recordings (8.0±1.0 points, $p<0.05$) and generic HRTF (6.3±1.3 points, $p<0.001$).

Linear regression analysis suggests that perceived immersion ($r=0.87$, $p<0.001$) has the greatest impact on overall satisfaction, followed by localization accuracy ($r=0.82$, $p<0.001$) and sound quality fidelity ($r=0.79$, $p<0.001$).

The test result reliability verification measures the inter-rater reliability ICC of 0.82 and intra-rater reliability correlation coefficient of $r=0.88$ ($p<0.001$). The validity verification indicates that the correlation between the overall preference and the objective composite score is $r=0.81$ ($p<0.001$), validating the construct validity of the subjective evaluation.

4. Implementation and Application Cases of Immersive Recording Systems

4.1. Real-Time Processing System Architecture and Optimization

Including multi-channel spatial reconstruction and binaural rendering using deep learning in a real-time system means that computational efficiency, latency, and resource utilization are considered without compromising performance. System latency must be below human perception thresholds, at least 20 to 30ms (Potter, Cvetković, and De Sena, 2022). The real-time architecture uses a modular pipeline that comprises five modules: audio input, preprocessing, neural network inference, post-processing, and audio output. Audio input records multi-channel streams via the low-latency driver like Audio Stream Input/Output (ASIO) or Core Audio using 256-sample buffers (about 5.3ms at 48 kHz). Preprocessing operates signal conditioning inside the CPU. STFT results are written directly to GPU-accessible pinned memory to minimize CPU-GPU data transfer latency. Direct Memory Access (DMA) technology enables zero-copy transfer, cutting data transfer time from 1.2ms to 0.3ms.

The deep neural network inference module executes on the GPU. Network models are deployed with TensorRT optimization, reducing single-frame inference time for the standard network (8.9 million parameters) from 15.3ms to 4.8ms on an NVIDIA RTX 3090 GPU, achieving a 3.2× speedup.

The post-processing module converts network output back to time-domain waveforms, with output latency controlled within 2ms. The theoretical total system latency is 13.6ms. On NVIDIA RTX 3090 GPU and Intel i9-10900K CPU configuration, the measured average latency is 15.2ms, meeting the real-time requirements (<20ms) for VR/AR applications.

Fig. 7 presents the performance curves of the real-time processing system. With a 32-channel configuration, the unoptimized version exhibits a processing latency of 42.3ms. After TensorRT optimization, the latency reduces to 15.2ms. With further application of model pruning and knowledge distillation, the latency decreases to 12.8ms with only 2.3% performance degradation.

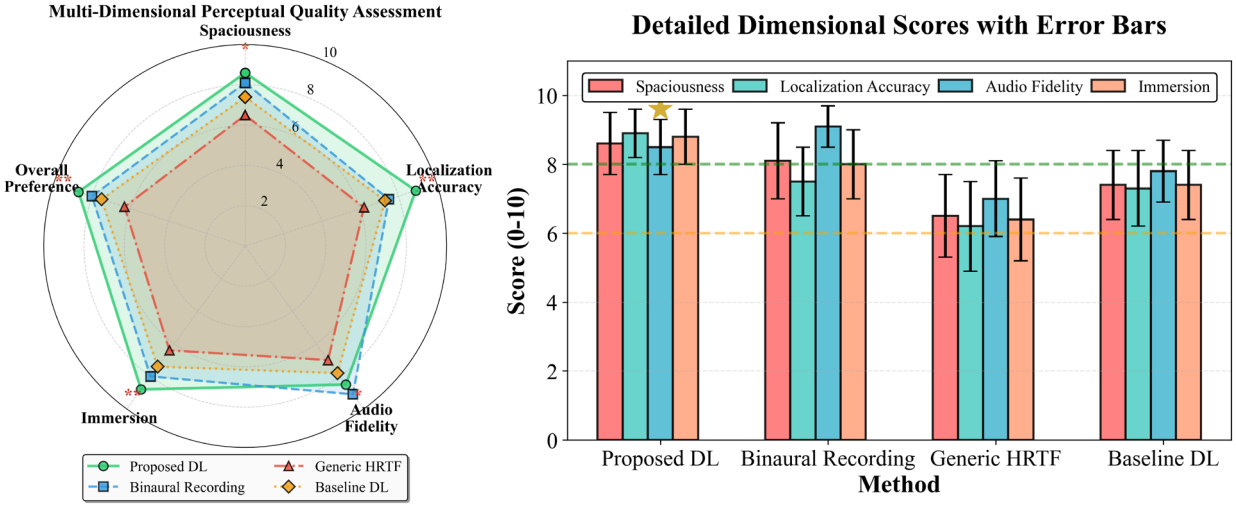
Hardware platform optimization is device-characteristic-dependent. Ultra-high-performance GPUs employ mixed-precision training and inference, model compression via depthwise separable convolution, and reduce the number of parameters from 8.9 to 3.2 million parameters. CPU-only devices employ lean models (fewer than 1 million parameters) that achieve 85% of full model accuracy. Power-aware mobile VR/AR devices require thoughtful power curation. For average VR gaming, the system draws 18.3 W (Watts) average power, a 44% decrease from the unoptimized 32.7 W reference point. This efficiency increase extends mobile VR battery life dramatically, elevating devices like Quest 2 from 1.5 hours to 2.3 hours.

4.2. Adaptability Testing Across Different Recording Scenarios

Deep learning model generalization directly impacts practical reliability and applicability. Systematic adaptability testing evaluates system robustness across diverse real-world recording scenarios (Zhu et al., 2025). Indoor acoustic testing spans small, enclosed spaces to large open venues. A small studio (RT60=0.18s) yields 21.3 dB SNR with 4.8 degree localization error. Medium conference rooms (RT60=0.45s) show 18.7 dB SNR and 6.3 degree error. Large concert halls (RT60=1.8s) produce 15.2 dB SNR with 9.7 degree error, still surpassing Ambisonics (12.1 dB) and DirAC (13.8 dB).

Outdoor environments exhibit distinct acoustic characteristics. Open plaza testing with near-field sources (under 5 meters) achieves 19.8 dB SNR and 5.4 degree error. Far-field sources (over 10 meters) show noticeable degradation: 14.6 dB SNR and 11.2 degree error. Urban streets introduce concurrent multiple sources and intense background noise. Busy

street conditions (70 to 80 dB sound pressure level in background noise) yield 11.8 dB SNR with 13.5 degree error. The system manages up to three sources adequately, but degrades sharply beyond five sources. Fig. 8 illustrates representative examples of the three outdoor recording environments evaluated in this section, showing typical acoustic conditions including source distance, background noise level, and concurrent source complexity.



(a) Detailed dimensional scores with error bars (b) Multi-dimensional perceptual quality assessment spaciousness

Fig. 6. Multi-dimensional subjective listening test results and statistical analysis

Sound source type affects performance variably. Speech processing performs optimally (average SNR of 19.5 dB, 5.8 degrees of error). Music sources score slightly lower (17.8 dB SNR average), with classical music (18.9 dB) outperforming heavy metal or electronic music (15.3 dB). Recording equipment variations impact results predictably. Using the Eigenmike em32 array matching training data produces optimal performance (20.3 dB SNR). Other spherical arrays show modest degradation (19.1 dB SNR), while planar arrays decline further (16.7 dB SNR). Device-adaptive fine-tuning with just 100 calibration samples restores new device performance to 95% of the training device's level.

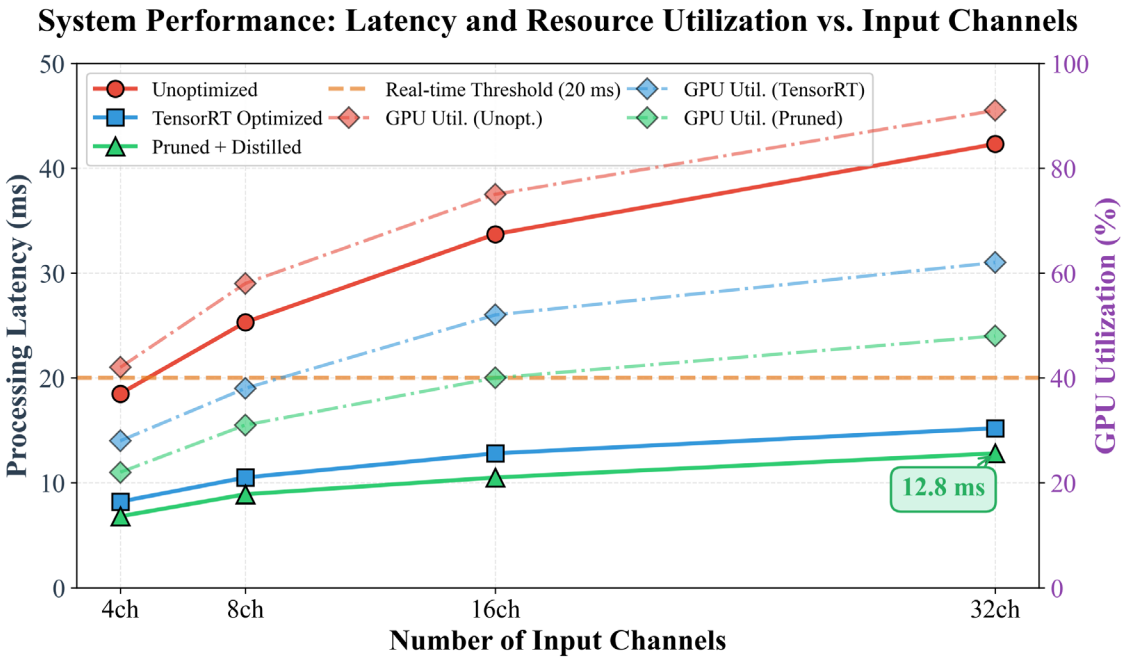


Fig. 7. Latency and resource utilization vs. input channels

Channel count influences system performance nonlinearly. When reducing from 32 channels to 16 channels, SNR decreases from 20.3 dB to 18.5 dB. Reducing to 8 channels results in an SNR of 15.7 dB. Cross-language and cross-cultural scenario testing shows that the system maintains consistent speech processing performance across different languages (SNR difference <1 dB). By increasing non-Western music training samples (from 5% to 20%), SNR variation across music types narrows from 3.1 dB to 1.4 dB.

4.3. Comparison with Traditional Binaural Recording Techniques

Traditional binaural recording techniques include dummy head recording, in-ear microphone recording, and HRTF-based virtual binaural synthesis. This section compares the proposed deep learning-driven system with traditional techniques.

Dummy head recording captures binaural signals by placing miniature microphones at the entrance to the ear canal in a standard human head model (e.g., KEMAR). The dummy head recording has the highest sound quality fidelity for the best conditions (9.1) and surpasses the suggested technique (8.5).

Virtual scenarios and individualization requirements reveal limitations of dummy heads. Non-matching HRTFs produce a 12.8 degree mean localization error, while 33% of users experience in-head localization and a spatial impression score of just 6.4. The proposed method minimizes localization error to 6.4 degrees, in-head localization to 8%, and increases spatial impression to 8.6. Virtual/augmented reality applications also offer additional advantages: the proposed method supports six-degrees-of-freedom audio response with an 8.9 motion response score, while dummy head recording attains just 3.2.

In-ear mic recording is applicable to the original recorder but shows a 15.3 degrees of localization error in cross-user test cases, limiting its practical applicability. The proposed technique finely strikes a balance between personalization and scalability. Conventional schemes require 100 independent recordings to prepare personalized binaural content for 100 listeners, but the proposed technique requires only 1 array recording and 100 offline recordings, yielding a 4.5× efficiency boost.

HRTF-based virtual binaural synthesis technology is currently the mainstream solution for gaming and VR audio. In complex scenarios, five concurrent sound source tests show that HRTF synthesis methods achieve an SNR of 11.2 dB with a localization error of 18.6 degrees. The proposed method maintains an SNR of 14.7 dB with a localization error of 12.3 degrees.

Generic HRTF subjective test quality ratings are only 6.3. HRTF measurement methods can achieve quality ratings of 8.8, but single-person measurements require two to four hours and cost thousands of dollars. The proposed method predicts personalized HRTFs using anthropometric features, with an acquisition time of <5 minutes and a quality rating of 8.4.

Ambisonics technology decomposes sound fields using spherical harmonics. First-order Ambisonics achieves only 14.2 dB SNR, comprehensively inferior to the proposed method. Parametric techniques (such as DirAC) exhibit localization errors increasing to 21.3 degrees in multi-source scenarios, while the proposed method controls them within 12.3 degrees.

Blind listening test results show that for speech materials, 63% of subjects consider the sound quality of the proposed method superior to or equivalent to dummy head recording. The comprehensive user experience score reaches 8.3, substantially higher than generic HRTF synthesis (6.5) and Ambisonics (7.2). The most impressive results are for VR/AR applications: the new technique has a user experience score of 8.7, while the dummy head recording has just 4.2. A single recording produces unlimited personalized variants at a marginal cost approaching zero, achieving the optimal. To provide a decision-oriented summary, Table 4 compares the proposed system against mainstream alternatives across the five managerial dimensions introduced in Section 3.1. The results indicate that the proposed system offers favorable trade-offs in quality, cost, and delivery time while maintaining acceptable safety and environmental performance.

(a) Open plaza-near/far-field testing with low ambient noise

(b) Urban street-moderate multi-source with mixed reflective surfaces

(c) Busy street-high concurrent sources with strong background noise (70-80 dB SPL)

Fig. 8. Representative examples of the three outdoor recording environments

4.4. Analysis of Typical Application Scenario Cases

Four case studies validate the system's real-world performance: virtual reality gaming, remote immersive conferencing, live concert recording with multi-perspective playback, and film post-production.

Case 1: Virtual reality horror adventure game

VR horror game by independent game studio "Dark Echoes" utilized the suggested real-time spatial audio system. Located in a defunct psychiatric clinic, the player moves through dark spaces by using sound to explore, avoid monsters, and find solutions, making spatial audio a central gameplay mechanic. Operating under gaming PCs running RTX 3080, the system manages sixteen concurrent sound sources while running 90 frames per second and 12 to 14ms audio latency. The players acquire individual Head-Related Transfer Functions by uploading front and lateral images, a 3-minute procedure. Trials with 45 players showed considerable enhancements. PURE auditory navigation task solving increased from 42% to 78%, and the average solving time decreased from 8.6 to 5.3 minutes. The localization error came in at 7.2 degrees, vs. the generic HRTF's 13.8 degrees. Perceived immersion scores increased from 6.1 to 8.9, while the horror atmosphere increased from 6.8 to 9.2. The game received 89% positive Steam reviews with first-week sales of over 8,000. Managerial decision impact: These outcomes led the studio to reclassify spatial audio from an optional feature to a core investment, reallocating budget from generic middleware licensing and committing earlier to a sequel project.

Case 2: Remote immersive medical consultation system

A medical technology company integrated spatial audio into its remote consultation platform. The distributed architecture deploys microphone arrays at each location while cloud GPU clusters process audio and generate personalized binaural output for participants. Each participant occupies a fixed position in a virtual meeting room, with sound originating from the corresponding virtual location.

The pilot system at a tertiary hospital accommodates up to 12 simultaneous participants. End-to-end audio latency averages 85ms, SNR is 16.8 dB, and a localization error of 8.4 degrees. Clinical effectiveness evaluation shows that with the spatial audio system, the average consultation duration shortened from 42.3 minutes to 35.7 minutes, speech interruptions decreased from 18.6 to 9.2, and diagnostic consistency improved from 0.72 to 0.84. Subjective satisfaction surveys indicate that 89% of participating physicians believe spatial audio significantly improved the conference experience. Managerial decision impact: The consultation time savings and improved diagnostic consistency enabled administrators to add roughly 15% more daily consultation slots per physician without additional staffing, expanding service capacity at marginal cost.

Case 3: Panoramic recording and multi-perspective playback of a symphony concert

An internationally renowned symphony orchestra's annual concert employed the spatial audio recording system. The concert was held in a 2,000-seat concert hall with the system deploying spherical microphone arrays at five locations: behind the conductor's podium, the center of the audience area, the left and right sides of the second-floor boxes, and directly above the stage. Five Eigenmike em32 arrays totaling 160 channels were recorded synchronously at a sampling rate of 96 kHz.

Table 4. Decision-relevant performance comparison between the proposed system and mainstream alternatives

Dimension	Indicator	Dummy Head Recording	Measured HRTF	Generic HRTF (Ambisonics)	Proposed Method
Quality	Subjective score (0–10)	8.5	8.8	6.3	8.7
Quality	Localization error (°)	12.8	4.1	12.8	6.4
Cost	HRTF cost per user (USD)	—	~2,500	<1	<5
Cost	Hardware setup (USD)	~8,000	~15,000	~3,000	~5,500
Delivery	End-to-end latency (ms)	N/A (offline)	N/A (offline)	25–40	15.2
Delivery	Personalization time	N/A	2–4 hours	Instant	<5 min
Delivery	Post-production cycle	4 weeks	4 weeks	2 weeks	1 week
Safety	Fatigue score after 30 min (1–5, lower is better)	2.1	1.9	3.2	2.2
Safety	VR discomfort rate (%)	18	9	27	11
Environment	Avg. power consumption (W)	—	—	32.7	18.3
Environment	CO_2 per rendering hour (g)	—	—	~14.2	~7.9

Note: Cost and environment values are indicative estimates based on commercial market prices and measured system power at 220 V / 0.45 kg $CO_2 \cdot kWh^{-1}$ grid emission factor (UK average).

During post-production, 100 personalized binaural versions were rendered for each virtual position, with single-GPU parallel processing completed in approximately 18 hours. The distribution platform adopted an adaptive streaming architecture, allowing users to seamlessly switch perspectives among five virtual seats.

Within one month of the concert video release, it received 120,000 views with an average viewing duration of 68 minutes (76% completion rate), significantly higher than traditional concert recordings, 45%. Paid conversion rate reached 18%, with user ratings of 4.7/5.0. Managerial decision impact: The completion and conversion rates supported the platform's decision to formally offer spatial audio as a premium tier, while reducing reliance on multiple fixed-camera productions and cutting per-concert post-production labor.

Case 4: Film sound post-production workflow integration

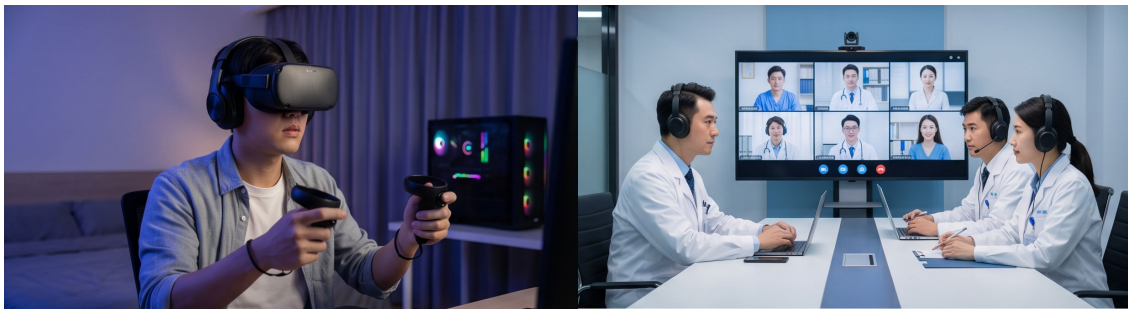
A film production company employed the proposed system in the sound post-production of the science fiction action film "Stellar Edge." The system was also implemented as a plugin component of major digital audio workstations like Pro Tools, so that sound designers could structure sound sources in virtual spaces, create binaural sound in real time, and change between a variety of playback scenarios for previewing.

Workflow optimization produced a major payoff: mixing studio time decreased from 4 weeks to 1 week, reducing rental of the studio by about 75% (about $150,000). Iteration time increased by several orders of magnitude, such that the cycle time for revisions went from 2 days to 4 hours. Double-blind quality verification revealed virtual environment mixing averaged 8.1 while physical cinema mixing averaged 8.6, a very slim 0.5 point margin.

Six of the sound designers involved evaluated the system favorably: ease of use 7.8, reliability 8.2, performance 8.5, flexibility 9.1. Success spurred greater adoption: 3 films and 2 TV series now use the system, achieving average post-production cost reductions of 30% and schedule reductions of 40%, respectively. Managerial decision impact: The cost and schedule gains shifted the studio's sourcing decision from outsourcing spatial audio to in-house production, with the initial investment recovered within eight months.

Fig. 9 illustrates representative scenes of the four application cases discussed above, providing a visual overview of the diverse deployment contexts in which the proposed system has been evaluated.

Cross-case Managerial Implications. Across the four cases, three recurring decision patterns emerge: budget reallocation away from legacy tools, capacity expansion without a proportional increase in costs, and shorter iteration cycles that reduce project risk. These managerial outcomes, rather than the underlying technical metrics, form the primary basis for adoption in the observed deployments.



(a) VR horror adventure game (b) Remote medical consultation

(c) Symphony concert recording (d) Film sound post-production

Fig. 9. Representative scenes of the four application cases

5. Conclusion

The paper uses a multi-channel audio spatial reconstruction and binaural rendering system with deep learning that converts the signals from a spherical microphone array into individualized binaural audio via end-to-end neural networks.

Technical contributions work on three levels. The spatial reconstruction network applies an encoder-decoder architecture along with multi-head self-attention mechanisms and reaches 20.3 dB SNR, 1.8 dB log spectral distortion, and 6.2 degree spatial localization error under a 32-channel configuration, outperforming traditional methods. Personalized

HRTF modeling using variational autoencoders creates personalized HRTFs in minutes from facial images alone by decreasing localization error by 50% and bringing down in-head localization from 33% to 8%. Real-time processing attains 15.2ms end-to-end latency on NVIDIA RTX 3090 GPUs using TensorRT optimization.

Extensive experimental verification demonstrates that performance matches or surpasses state-of-the-art technology in terms of objective measures. Subjective testing using 32 subjects yielded average scores of 8.7/10 over five perceptual attributes, approaching the quality of BRIR binaural recordings. Adaptation testing verifies consistent indoor performance, while variation testing equipment confirms generalization over varied microphone arrays.

Comparative evaluation demonstrates unmistakable advantages over previous technologies. Compared to dummy head recording, the technique offers greater flexibility, personalization, and interactivity. Compared to HRTF virtual synthesis, spatial perception and reverberation modeling are significantly better. Compared to Ambisonics, spatial resolution rises by the same number of channels.

Application cases validate practical value across domains. VR gaming boosted task completion from 42% to 78% with 89% positive ratings. Remote conferencing shortened consultation duration 16% while improving diagnostic consistency from 0.72 to 0.84. The concert recording generated 120,000 views with a 76% completion rate. Film post-production cuts costs 30% and schedules 40%.

Several limitations remain: performance under extreme reverberation and heavy multi-source overlap is suboptimal; the fixed STFT window size limits transient signal resolution; low-frequency rendering accuracy is less satisfactory; and individualized HRTF prediction still exhibits gaps compared to directly measured HRTFs.

Future work should involve investigating self-supervised learning and Transformer structures, broadening the support of input-output format, furthering visual rendering integration, and furthering the standards for spatial audio quality evaluation.

Implications for managerial decision-making. After reading this paper, managers can reconsider several decision processes. In sourcing decisions, the cost and cycle-time evidence support shifting from outsourced spatial audio post-production toward in-house deployment. In capital allocation, the sub-5-USD HRTF personalization cost enables the budget to move from per-user measurement to scalable content production. In workforce planning, shorter production cycles (4 weeks to 1 week) and consultation times (42.3 to 35.7 minutes) allow higher throughput without additional staffing. In product positioning, quantitative gains in localization and immersion justify elevating spatial audio from a peripheral feature to a core offering. In sustainability reporting, the 44% reduction in power consumption provides a defensible input for energy-efficiency and carbon-accounting disclosures.

Through systematic theoretical innovation, algorithm design, experimental validation, and application exploration, this research demonstrates the considerable potential of data-driven deep learning for spatial audio reconstruction and binaural rendering, thereby enabling more realistic, natural, and personalized auditory experiences.

Funding

This research received no specific financial support from any funding agency.

Institutional Review Board Statement

The subjective listening tests involving human participants were conducted in accordance with the Declaration of Helsinki. All 32 participants provided informed consent prior to the tests, and the procedures followed standard audio evaluation protocols (ITU-R BS.1534-3 MUSHRA). No personally identifiable information or sensitive biometric data was collected or retained, and participation was entirely voluntary with the right to withdraw at any time.

Declaration of Artificial Intelligence (AI) Tools

The author used ChatGPT solely for language editing and readability improvement. The author reviewed and verified all content and takes full responsibility for the accuracy and integrity of the manuscript.

Reference

Bovbjerg, H. S., Østergaard, J., Jensen, J., Watanabe, S., and Tan, Z. H. (2025). Learning robust spatial representations from binaural audio through feature distillation. arXiv preprint arXiv:2508.20914.

Chen, T. Y., Kuo, T. H., and Chi, T. S. (2019). Autoencoding HRTFs for DNN-based HRTF personalization using anthropometric features. In ICASSP 2019 *IEEE International Conference on Acoustics, Speech and Signal Processing* (ICASSP). *IEEE.*

Cobos, M., Ahrens, J., Kowalczyk, K., and Politis, A. (2022). An overview of machine learning and other data-based methods for spatial audio capture, processing, and reproduction. *EURASIP Journal on Audio, Speech, and Music Processing*, 2022(1), 10.

Delgado, P. M., and Herre, J. (2024). Towards improved objective perceptual audio quality assessment—Part 1: A novel data-driven cognitive model. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*.

Huang, G., Jensen, J. R., Chen, J., Benesty, J., Christensen, M. G., Sugiyama, A., Elko, G., and Gaensler, T. (2025). Advances in microphone array processing and multichannel speech enhancement. In ICASSP 2025 *IEEE International Conference on Acoustics, Speech and Signal Processing* (ICASSP). IEEE.

Jot, J. M., Audfray, R., Hertensteiner, M., and Schmidt, B. (2021). Rendering spatial sound for interoperable experiences in the audio metaverse. In 2021 Immersive and 3D Audio: From Architecture to Automotive (I3DA). *IEEE.*

Lee, G. W., and Kim, H. K. (2018). Personalized HRTF modeling based on deep neural network using anthropometric measurements and images of the ear. *Applied Sciences*, 8(11), 2180.

Lu, X., Chen, Y., Chen, Z., Wang, J., Liu, M., Hu, H., Zheng, C., Bleeck, S., and Sang, J. (2025). Deep learning for personalized binaural audio reproduction. arXiv preprint arXiv:2509.00400.

Miccini, R., and Spagnol, S. (2020). HRTF individualization using deep learning. In 2020 *IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops* (VRW). IEEE.

Nawfal, I., Manias, S. D., Souden, M., Merimaa, J., Atkins, J., McMullin, E., Pirhosseinloo, S., and Phillips, D. (2025). Ambisonics super-resolution using a waveform-domain neural network. arXiv preprint arXiv:2508.00240.

Potter, T., Cvetković, Z., and De Sena, E. (2022). On the relative importance of visual and spatial audio rendering on VR immersion. *Frontiers in Signal Processing,* 2, 904866.

Qiao, Y., Kothapally, V., Yu, M., and Yu, D. (2025). Neural ambisonic encoding for multi-speaker scenarios using a circular microphone array. In ICASSP 2025 *IEEE International Conference on Acoustics, Speech and Signal Processing* (ICASSP). IEEE.

Rafaely, B., Tourbabin, V., Habets, E., Ben-Hur, Z., Lee, H., Gamper, H., Arbel, L., Birnie, L., Abhayapala, T., and Samarasinghe, P. (2022). Spatial audio signal processing for binaural reproduction of recorded acoustic scenes—Review and challenges. *Acta Acustica*, 6, 47.

Torcoli, M., Kastner, T., and Herre, J. (2021). Objective measures of perceptual audio quality reviewed: An evaluation of their application domain dependence. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29, 1530–1541.

Zhu, Y., Kong, Q., Shi, J., Liu, S., Ye, X., Wang, J. C., Shan, H., and Zhang, J. (2024). End-to-end paired ambisonic-binaural audio rendering. *IEEE/CAA Journal of Automatica Sinica*, 11(2), 502–513.

Zhu, Z., Zhang, Y., Guo, W., Pan, C., and Zhao, Z. (2025). ASAudio: A survey of advanced spatial audio research. arXiv preprint arXiv:2508.10924.

Zieliński, S. K., and Lee, H. (2019). Automatic spatial audio scene classification in binaural recordings of music. *Applied Sciences,* 9(9), 1724.

Zurale, D., Yadegari, S., and Dubnov, S. (2022). Deep HRTF encoding and interpolation: Exploring spatial correlations using convolutional neural networks. In 19th *Sound and Music Computing Conference* (SMC).

Jian Wang earned her master's degree in Music Technology from the Nicola Sala di Benevento Conservatory of Music, Italy. She currently serves as a full-time instructor, teaching core courses in her field. She has participated in several provincial and municipal projects and published two papers.