



Audio Engineering Society

Convention Paper 10244

Presented at the AES 159th Convention
2025 October 23–25, Long Beach, CA, USA

This paper was peer-reviewed as a complete manuscript for presentation at this convention. This paper is available in the AES E-Library (<http://www.aes.org/e-lib>), all rights reserved. Reproduction of this paper, or any portion thereof, is not permitted without direct permission from the Journal of the Audio Engineering Society.

Perceptual Quality-Preserving Neural Audio Compression for Low-Bandwidth VR

Sanjay Majumder

Style tree

Correspondence should be addressed to Sanjay Majumder (sanjay@styletree.me)

ABSTRACT

As immersive technologies such as Virtual Reality (VR) and Augmented Reality (AR) become increasingly prevalent, the demand for efficient, perceptually optimized audio compression has grown rapidly. Traditional audio codecs prioritize signal fidelity over perceptual relevance and often fail to preserve critical spatial cues under low-bitrate constraints. In this paper, we propose a novel perceptual neural audio codec tailored for low-bandwidth immersive applications, with an emphasis on preserving both auditory salience and spatial integrity. Our approach combines psychoacoustic feature extraction with a deep encoder–decoder architecture, conditioned on spatial metadata and trained using perceptual and spatially aware loss functions. The codec is evaluated across multiple datasets, including speech and ambient scenes, demonstrating superior performance over existing classical and neural codecs in terms of both objective metrics (e.g., SNR, LSD, ViSQO [1]) and subjective Mean Opinion Scores (MOS). Moreover, the lightweight model architecture enables real-time decoding on mobile hardware, making it suitable for deployment in VR/AR headsets, mobile devices, and low-power audio systems. This work establishes a foundation for scalable, high-quality spatial audio delivery under severe bandwidth constraints, advancing the state-of-the-art in perceptual audio coding for immersive media.

1 Introduction

The proliferation of immersive media platforms, particularly in Virtual Reality (VR) and Augmented Reality (AR), has significantly elevated user expectations for high-quality audio experiences. Spatial audio, which simulates the perception of sound in three dimensions, plays a critical role in establishing a sense of presence and realism in virtual environments. However, delivering high-fidelity spatial audio in bandwidth-constrained settings remains a major challenge, especially for wireless VR systems and mobile devices. Traditional audio codecs, while effective for general-purpose audio compression, are not specifically designed to preserve

psychoacoustic and spatial cues critical to immersion. This often results in perceptual artifacts, degraded localization accuracy, and a reduced sense of presence when used in low-bitrate scenarios.

Recent advances in deep learning-based audio compression have demonstrated impressive gains in efficiency and quality by leveraging neural networks to model complex audio structures. However, most of these models prioritize waveform reconstruction fidelity and overlook the psychoacoustic and spatial characteristics essential for immersive content. Furthermore, the unique demands of VR audio, including real-time performance, head-tracked spatial consistency, and minimal percep-

tual distortion, necessitate a fundamentally different approach to codec design.

This paper introduces a perceptually motivated neural audio codec suited for immersive applications in low-bandwidth VR environments. The proposed system employs a feature-aware encoding strategy that prioritizes psychoacoustically salient components such as tonality, masking thresholds, and spatial localization cues (e.g., interaural time and level differences). A neural encoder-decoder architecture learns a compressed representation optimized not only for waveform fidelity, but also for perceptual integrity and spatial realism. An adaptive bitrate allocation mechanism further ensures efficient bandwidth utilization based on audio complexity and scene dynamics.

By integrating perceptual modeling with spatially aware neural compression, the proposed approach aims to preserve the immersive experience of the listener while operating within strict bandwidth constraints. This work contributes a novel codec design, a rigorous evaluation framework combining objective metrics and VR-based listening tests, and experimental validation demonstrating competitive performance against state-of-the-art audio codecs at significantly lower bitrates. In this paper, we use the term ‘spatial audio’ to refer specifically to *binaural (two-channel) headphone playback*, that is, HRTF-convolved renderings or stereo signals with strong interaural differences. Our system is therefore designed to preserve binaural localization cues (ITD, ILD, IACC) under bandwidth constraints, rather than to encode native multichannel formats (e.g., 5.1, HOA, or object-based audio).

2 Related Work

Audio compression has been a long-standing area of research, with traditional codecs such as MP3, Advanced Audio Coding (AAC) and Opus [2] offering mature, low-bitrate solutions for general audio content. Although these codecs incorporate psychoacoustic models to remove perceptually irrelevant components, they are primarily optimized for stereo or mono playback and do not preserve the spatial audio cues necessary for immersive virtual environments. For instance, Opus is widely used for voice over IP and streaming applications due to its low latency and reasonable quality, but it lacks mechanisms to handle interaural differences or encode head-tracked spatial information accurately.

To address some of these limitations, spatial audio coding standards such as MPEG-H and Ambisonics have been developed, allowing multi-channel representations and head-related transfer function (HRTF) rendering. However, these methods typically rely on large bandwidths or complex rendering chains and still require efficient compression schemes to be viable in mobile or real-time applications. Moreover, their reliance on handcrafted spatial models limits adaptability across diverse listening environments and user-specific head-related transfer functions.

In recent years, neural audio compression methods have emerged, leveraging deep learning to outperform traditional codecs in both objective and subjective evaluations. Works such as SoundStream [3], Encodec [4], and HiFiCodec [5] demonstrate end-to-end learning of audio representations using convolutional or transformer-based encoder-decoder architectures with vector quantization and adversarial training. Although these approaches significantly improve waveform reconstruction quality, they generally focus on speech or general music signals and do not explicitly model perceptual relevance or spatial fidelity.

Perceptual modeling in neural audio has been explored through loss functions based on psychoacoustic features, including auditory filterbanks, masking curves, and differentiable perceptual metrics such as ViSQO-LAudio [1] or Perceptual Evaluation of Audio Quality (PEAQ) [6]. However, integration of such models into a rate-controlled compression pipeline optimized for immersive audio remains limited. Furthermore, most neural codecs ignore critical spatial attributes such as interaural time difference (ITD), interaural level difference (ILD), and spatial coherence, which are essential for VR and AR experiences.

Building upon these foundations, this work proposes a neural codec that explicitly incorporates perceptual and spatial modeling into its compression framework. By unifying deep audio representation learning with psychoacoustic feature extraction and spatial cue preservation, we aim to fill a critical gap in neural audio coding for bandwidth-limited immersive environments.

3 System Overview

The proposed system is a perceptually driven, end-to-end neural audio codec designed for immersive audio compression in bandwidth-constrained environments

such as virtual and augmented reality. The architecture consists of three main components: an encoder that transforms raw binaural audio (two-channel, HRTF-convolved signals intended for headphone playback) into a compact latent representation, a quantizer that enforces bitrate control through discrete tokenization of the latent space, and a decoder that reconstructs the binaural signal while preserving psychoacoustic and spatial fidelity. In this work, we focus specifically on binaural content rather than native multichannel or object-based spatial audio formats, and therefore the scope of “spatial audio” refers to the preservation of interaural cues (ITD, ILD, IACC) under low-bitrate conditions.

The encoder utilizes a series of 1-D temporal convolutional layers to extract hierarchical audio features while preserving time-frequency resolution and spatial coherence. This feature extractor is enhanced with perceptual modules that model tonality, masking thresholds, and spatial cues such as interaural time and level differences. The output is passed through a vector quantization module or an entropy-constrained bottleneck that ensures bitrate control suitable for streaming in low-bandwidth settings.

The decoder mirrors the encoder structure and reconstructs the binaural audio signal using Conv1DTranspose layers with a focus on spatial localization accuracy and perceptual integrity. In addition to waveform reconstruction loss, the training objective includes perceptual and spatial losses to ensure high-quality immersive audio. An adaptive bitrate allocation module monitors the content complexity and redistributes bits accordingly, ensuring optimal resource usage across varying acoustic scenes. Figure 1 provides an overview of the complete codec architecture.

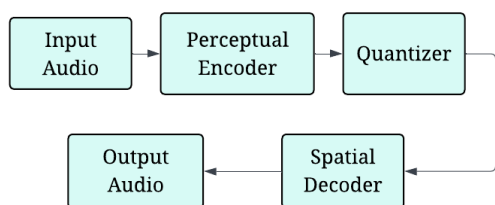


Fig. 1: Block diagram of the proposed audio codec.

4 Psychoacoustic Feature Extraction

To achieve high perceptual quality at very low bitrates, it is essential to identify and prioritize the audio components most relevant to human auditory perception. The proposed system integrates a psychoacoustic feature extraction module into the encoder pipeline to guide the compression process using auditory salience rather than purely signal-based heuristics. This module extracts a set of psychoacoustic features inspired by models of the human auditory system, including tonality, masking thresholds, critical band energy distribution, and temporal masking patterns.

First, the input audio is transformed into a time-frequency representation using a short-time Fourier transform (STFT) that approximates the critical bands of the Bark or ERB scale. Tonality is computed by analyzing the spectral flatness and harmonic structure across frames, which helps the encoder distinguish between noise-like and tone-like components. This information is used to prioritize tonal regions during compression, as tonal artifacts are more perceptible to human listeners.

Spectral masking models are applied simultaneously to estimate the masking threshold for each time-frequency bin based on surrounding energy and psychoacoustic masking curves. These thresholds are then used to generate a perceptual importance map that weighs the perceptual relevance of each region, allowing the encoder to discard or coarsely quantize masked components with minimal impact on perceived quality. Temporal masking is also modeled by considering forward and backward masking effects, allowing more aggressive compression of transient regions that are perceptually overshadowed by preceding or succeeding high-energy events.

In addition, for binaural and spatial audio inputs, spatial psychoacoustic features such as interaural time differences (ITD), interaural level differences (ILD), and coherence between channels are extracted to preserve accurate spatial localization. These spatial features are fused with the spectral and temporal psychoacoustic map to inform the encoder about perceptually salient spatial cues that must be preserved across the quantization and decoding stages.

By incorporating these psychoacoustic features directly into latent space encoding and loss functions, the system ensures that compression decisions are aligned

with perceptual relevance rather than naive signal metrics. This leads to higher-quality audio reconstructions that maintain both timbral fidelity and immersive spatial characteristics, even at low bitrates.

5 Neural Codec Architecture

The neural codec architecture proposed in this work is designed to deliver efficient, high-quality compression of binaural spatial audio, with specific optimizations for bandwidth-constrained immersive applications such as virtual reality (VR) and augmented reality (AR). The architecture follows an encoder–quantizer–decoder paradigm, but with significant augmentations to integrate psychoacoustic modeling and spatial awareness directly into the learning and inference process. Each component is carefully engineered to capture relevant auditory and spatial cues while maintaining computational efficiency suitable for deployment on mobile or real-time systems.

5.1 Perceptual Encoder

The encoder consists of a deep stack of convolutional layers that transform the raw multi-channel audio waveform into a compressed latent representation. For binaural or multichannel input, the encoder operates on each channel jointly, using shared weights with cross-channel feature fusion mechanisms. These fusion modules, implemented via grouped or depthwise separable convolutions followed by attention-based spatial alignment layers, explicitly capture interaural dependencies such as interaural time differences (ITD), interaural level differences (ILD), and phase coherence, features critical to maintaining spatial immersion.

To enhance perceptual selectivity, the encoder incorporates psychoacoustic guidance modules that apply importance weighting to feature maps based on externally computed saliency maps. These maps highlight time-frequency regions of perceptual significance, such as tonal components, unmasked harmonics, and transient attacks, using learned masks derived from a differentiable auditory model. This weighting mechanism biases the encoder to allocate more representational capacity to salient perceptual content, enabling more efficient compression of less relevant regions without noticeable degradation in perceived quality.

5.2 Quantization and Bitrate Control

Following the encoder, the latent representation is passed through a quantization bottleneck designed to enforce strict bitrate constraints. We adopt a vector-quantized autoencoder (VQ-VAE) framework with a learned codebook of size $K = 1024$ and embedding dimension $D = 128$, which maps continuous embeddings to discrete token indices. To avoid codebook collapse and promote usage diversity, we update the codebook using the exponential moving average (EMA) scheme introduced by Oord et al.[7], and apply a commitment loss with coefficient $\beta = 0.25$. This mechanism ensures stable utilization of the full codebook without requiring explicit clustering steps.

To support variable bitrate modes, the system includes a content-aware bitrate controller that dynamically adjusts the number of active codebooks or quantization resolution based on the estimated perceptual complexity of each input segment. This is achieved through an auxiliary bitrate prediction network that conditions the saliency map and encoder activations to estimate the entropy levels and adjust the quantization granularity accordingly. This adaptive control enables graceful degradation under extreme bandwidth constraints while preserving high fidelity in perceptually critical scenes.

5.3 Spatial Decoder

The decoder mirrors the encoder in structure, but operates in reverse, reconstructing the waveform from the quantized latent tokens. To ensure faithful reproduction of spatial cues, the decoder includes spatial coherence regularization layers that enforce consistency in ITD and ILD across the output channels. Additionally, phase reconstruction modules are employed to restore spatial phase alignment using group-delay-informed inverse filtering and differentiable phase unwrapping layers. These modules are particularly important for preserving the localization cues that underlie spatial audio perception.

The final stage of the decoder applies a transposed convolutional or learned synthesis filterbank to resample the compressed representation back to the waveform domain. This synthesis filterbank is trained jointly with the encoder and decoder, enabling end-to-end optimization. For computational efficiency, the decoder architecture is kept lightweight, with depth and complexity dynamically scaled based on target latency and compute budgets.

5.4 Training Paradigm

The entire codec is trained end-to-end using a multi-objective loss function that balances waveform fidelity, perceptual similarity, and spatial accuracy. In addition to traditional losses such as mean squared error (MSE) and spectral convergence loss in the time-frequency domain, the training objective incorporates perceptual losses based on auditory models (e.g., differentiable ViSQOLAudio) and learned perceptual similarity networks. To preserve spatial fidelity, a novel spatial loss term is introduced, penalizing differences in binaural cues between original and reconstructed signals, such as ITD, ILD, and interaural cross-correlation (IACC).

The training dataset consists of a diverse collection of spatial audio content that includes speech, music, environmental ambiences, and synthetic spatial soundscapes. Augmentation techniques such as HRTF variation, reverberation, and head rotation simulation are applied to enhance generalization. The model is trained using Adam Optimizer with warm-up learning rate scheduling, and quantization-aware training is employed from the early stages to ensure convergence under bitrate constraints.

This neural codec architecture represents a principled fusion of perceptual audio coding, spatial signal processing, and modern deep learning. It delivers a practical and scalable solution for perceptually optimized spatial audio compression, capable of operating under strict bandwidth and latency constraints while maintaining immersive quality across diverse listening environments.

6 Loss Functions and Training Procedure

To ensure that the proposed neural audio codec achieves perceptually transparent compression while preserving spatial cues, a comprehensive multi-objective loss function is formulated that jointly optimizes for waveform fidelity, perceptual similarity, and spatial accuracy. Each component of the loss is grounded in established psychoacoustic and spatial hearing principles, as well as recent advancements in differentiable perceptual models. These loss terms are combined with task-specific weights, enabling end-to-end learning that aligns model performance with human auditory perception.

6.1 Waveform and Spectral Losses

The most fundamental component of the loss function is the waveform reconstruction error, computed using the mean squared error (MSE) between the original and reconstructed waveforms:

$$\mathcal{L}_{\text{MSE}} = \frac{1}{T} \sum_{t=1}^T (x_t - \hat{x}_t)^2 \quad (1)$$

where T is the number of audio samples, x_t and $\hat{x}_t$ are the original and reconstructed signals at time t . Although MSE is simple and effective, it does not always correlate with perceived audio quality.

To address this limitation, we include spectral losses that operate in the time-frequency domain. Specifically, we compute the magnitude STFT loss, which measures the difference between the log-magnitude spectrograms of the original and reconstructed audio:

$$\mathcal{L}_{\text{STFT}} = \frac{1}{FT} \sum_{f=1}^F \sum_{t=1}^T \left[\log (|\text{STFT}(x)_{f,t}| + \epsilon) - \log (|\text{STFT}(\hat{x})_{f,t}| + \epsilon) \right]^2 \quad (2)$$

where the STFT is computed on log-magnitude spectra using a Hann window of 1024 samples, an FFT length of 1024, and a hop size of 256 samples (75% overlap) at a sampling rate of 16 kHz. We set $\epsilon = 10^{-9}$ to avoid numerical instability in the logarithm.

This spectral loss encourages accurate reconstruction of frequency content, including transients and harmonics, and is particularly useful for preserving the character of complex soundscapes.

6.2 Perceptual Losses

Beyond signal fidelity, the codec is guided by perceptual losses designed to approximate human listening preferences. We use a differentiable perceptual loss based on ViSQOLAudio, a neural model trained to predict perceptual similarity scores:

$$\mathcal{L}_{\text{Perceptual}} = 1 - \text{ViSQOL}(x, \hat{x}) \quad (3)$$

This loss penalizes perceptual artifacts such as timbral degradation, ringing, and smearing that might not be captured by signal-based losses.

In addition, a masking-weighted perceptual loss inspired by psychoacoustic models is integrated. This loss applies higher penalties to regions of the signal that are above the estimated masking threshold and lower penalties to masked regions, allowing the codec to focus on perceptually relevant content:

$$\mathcal{L}_{\text{Mask}} = \frac{1}{FT} \sum_{f=1}^F \sum_{t=1}^T w_{f,t} \left[|\text{STFT}(x)_{f,t}| - |\text{STFT}(\hat{x})_{f,t}| \right]^2 \quad (4)$$

Here, $w_{f,t}$ denotes the masking weight at time t and frequency bin f , defined as

$$w_{f,t} = \frac{\max(0, |\text{STFT}(x)_{f,t}| - \tau_{f,t})}{\sum_f \max(0, |\text{STFT}(x)_{f,t}| - \tau_{f,t}) + \varepsilon},$$

where $\tau_{f,t}$ is the psychoacoustic masking threshold and $\varepsilon = 10^{-9}$ avoids division by zero. This weighting ensures that supra-threshold (audible) frequency bins contribute proportionally to the loss.

6.3 Spatial Consistency Loss

To maintain spatial fidelity, critical in immersive audio, we introduce a spatial consistency loss that quantifies the deviation in spatial cues between the input and reconstructed binaural signals. The primary components include interaural time difference (ITD), interaural level difference (ILD), and interaural cross-correlation (IACC):

$$\begin{aligned} \mathcal{L}_{\text{Spatial}} = & \lambda_{\text{ITD}} \|\text{ITD}(x) - \text{ITD}(\hat{x})\|_2^2 \\ & + \lambda_{\text{ILD}} \|\text{ILD}(x) - \text{ILD}(\hat{x})\|_2^2 \\ & + \lambda_{\text{IACC}} \|\text{IACC}(x) - \text{IACC}(\hat{x})\|_2^2 \end{aligned} \quad (5)$$

Interaural time difference (ITD) was estimated as the peak of the generalized cross-correlation with phase transform (GCC-PHAT) computed over 20 ms windows (320 samples at 16 kHz) with 50% overlap. Reported ITD values are expressed in milliseconds by converting sample lags to time. Interaural level difference (ILD) was calculated as the mean left–right energy difference across one-third-octave bands with centre frequencies from 250 Hz to 8 kHz (250, 315, 400, ..., 8 000 Hz). Interaural cross-correlation (IACC) was computed as the maximum normalized cross-correlation within a

± 1 ms lag range (16 samples at 16 kHz), using 100 ms analysis windows and averaging results across time. This windowing provides a robust perceptual measure of binaural coherence. We set the weights to $\lambda_{\text{ITD}} = 1.0$, $\lambda_{\text{ILD}} = 0.5$, and $\lambda_{\text{IACC}} = 0.5$, balancing localization accuracy with perceptual quality.

6.4 Total Loss and Training Procedure

The total loss function is a weighted sum of the following terms:

$$\begin{aligned} \mathcal{L}_{\text{total}} = & \alpha \mathcal{L}_{\text{MSE}} + \beta \mathcal{L}_{\text{STFT}} + \gamma \mathcal{L}_{\text{Perceptual}} \\ & + \delta \mathcal{L}_{\text{Mask}} + \eta \mathcal{L}_{\text{Spatial}} \end{aligned} \quad (6)$$

where we set $\alpha = 1.0$, $\beta = 1.0$, $\gamma = 0.5$, $\delta = 0.5$, and $\eta = 0.5$ based on validation tuning.

Training is performed on a large-scale dataset of diverse spatial audio content, including speech, and environmental ambiences. Data augmentation techniques such as HRTF variation, random reverberation, and head movement simulation are employed to improve robustness. The model is trained using the Adam optimizer with cosine learning rate annealing and early stopping based on a validation perceptual loss. Gradient clipping and quantization-aware training (QAT) are applied from early epochs to stabilize training under bitrate constraints. The encoder, quantizer, and decoder are trained jointly in an end-to-end fashion, allowing the system to learn optimal representations and reconstruction strategies for high-quality and bandwidth-efficient audio compression.

This comprehensive loss framework ensures that the neural codec produces output that not only scores well on conventional metrics but also delivers subjectively superior and spatially immersive experiences in practical applications.

7 Adaptive Bit Allocation Strategy

To achieve optimal perceptual quality under stringent bitrate constraints, particularly in bandwidth-limited scenarios such as mobile virtual reality (VR) and augmented reality (AR), we implement a content-aware adaptive bit allocation strategy within our neural codec framework. Unlike fixed-rate compression schemes that uniformly distribute bits across the entire audio signal, our approach dynamically adjusts the bitrate allocation across temporal segments and frequency bands based on perceptual importance and spatial saliency.

All reported bitrates are *average transmitted bitrates* measured across the test set. Our codec operates in a variable bitrate (VBR) mode driven by the content-aware controller; when comparing to constant bitrate (CBR) waveform codecs we report matched average bitrates to ensure fair comparisons.

7.1 Perceptual Complexity Estimation

The foundation of the adaptive bit allocation mechanism is the real-time estimation of perceptual complexity. For each audio segment, we compute a psychoacoustic saliency map using a differentiable auditory front-end, which mimics aspects of the human auditory system such as critical band decomposition, temporal masking, and frequency masking. These saliency maps highlight time-frequency regions that are perceptually prominent, such as tonal components, transients, and speech formants, while suppressing masked or redundant components.

In parallel, we assess spatial complexity by analyzing interaural features such as interaural level difference (ILD) variance, interaural time difference (ITD) shifts, and spectral decorrelation across channels. High spatial complexity regions, such as moving sound sources or reverberant spatial scenes, are flagged for prioritized encoding to preserve localization cues and immersive realism.

7.2 Bitrate Controller Module

To enable adaptive allocation, we integrate a bitrate controller into the encoder pipeline. This module employs a lightweight Gated Recurrent Unit (GRU) based Recurrent Neural Network (RNN) that conditions on summary statistics of the latent activations (e.g., mean and variance across channels) and produces a dynamic bitrate allocation profile for each frame or segment. The output profile specifies allocation weights across frequency bands and temporal chunks, which are then used to control the resolution and number of active quantization bins in the vector quantizer.

To accommodate varying bandwidth requirements, the controller is trained with reinforcement learning using a reward function that balances perceptual fidelity (approximated by differentiable perceptual loss) against the actual bitrate consumed. During inference, the controller can be constrained to operate within a user-defined bitrate ceiling, ensuring practical deployment across heterogeneous devices and networks.

7.3 Scalable Layered Quantization

The quantization framework is extended to support scalable layered quantization, enabling a base layer and one or more enhancement layers. The base layer captures the perceptually dominant components of the audio, while the enhancement layers refine finer spectral details and spatial accuracy. Depending on the predicted complexity and target bitrate, the bitrate controller determines how many layers to activate. This layered design ensures graceful degradation in quality when bandwidth is limited and provides opportunities for progressive refinement when additional bits become available.

The entropy model used for layered latent compression is implemented as a masked convolutional hyperprior trained jointly with the decoder to minimize estimated bitrate $R = -\log_2 p(z)$ alongside reconstruction and perceptual losses. During deployment we use range coding for entropy coding of the latent indices.

7.4 Joint Training with Rate-Distortion Optimization

To ensure that the entire system learns to exploit perceptual redundancy effectively, we employ joint training of the encoder, quantizer, and bitrate controller using a rate-distortion optimization (RDO) loss:

$$\mathcal{L}_{\text{RDO}} = \mathcal{D}_{\text{perceptual}} + \lambda \cdot \mathcal{R} \quad (7)$$

where $\mathcal{D}_{\text{perceptual}}$ is a composite perceptual distortion term (including waveform, spectral, and spatial losses), $\mathcal{R}$ is the estimated bitrate of the quantized latent representation, and λ is a tunable rate-penalty coefficient that adjusts the trade-off between quality and bitrate. Bitrate estimation is performed using entropy models trained to predict the distribution of codebook usage, allowing efficient range coding during deployment.

The bitrate controller was trained using the REINFORCE algorithm with a learned baseline. The reward function used during controller training was

$$r = 0.7 \cdot Q_{\text{perceptual}} - 0.3 \cdot R \quad (8)$$

where $Q_{\text{perceptual}}$ is the differentiable perceptual quality proxy and R is the estimated bitrate. Training ran for 200k episodes with a horizon of 2 s per episode.

7.5 Runtime Bitrate Adaptation

Finally, to support real-time dynamic adaptation, especially in streaming or networked VR scenarios, the model exposes a control interface that adjusts λ and the controller’s target rate based on feedback from the network stack (e.g., congestion level, latency). This allows the codec to adapt continuously without retraining, offering a robust and flexible solution for real-world deployment in low-latency immersive systems.

The adaptive bit allocation strategy thus serves as a critical component in balancing fidelity, immersion, and efficiency. By tailoring bitrate usage to perceptual and spatial relevance, our codec ensures high-quality rendering of immersive audio experiences even under tight bandwidth budgets.

8 Evaluation Setup

To comprehensively evaluate the performance of the proposed perceptual neural audio codec under real-world constraints, a multi-faceted experimental setup that encompasses objective metrics, perceptual evaluations, and spatial fidelity analysis is designed. This section outlines the datasets, evaluation protocols, baseline models, and system configurations used in the assessment.

8.1 Datasets

To ensure robust generalization across both speech and environmental content, we used two primary datasets and spatialized them for binaural evaluation. The TUT Acoustic Scenes 2017 [8] dataset contains single-channel ambient recordings from diverse indoor and outdoor environments. We convolved these recordings with binaural room impulse responses (BRIRs) derived from the CIPIC HRTF database [9] to create spatialized versions suitable for evaluating scene consistency. Similarly, clean speech utterances from the LibriSpeech corpus [10] were similarly convolved with CIPIC-based BRIRs to simulate speech in realistic spatial settings.

All audio was resampled to 16 kHz to match the codec’s operating resolution and split into training (80%), validation (10%), and test (10%) sets. The CIPIC HRTF

database [9] provides individualized head-related transfer functions (HRTFs) at high spatial resolution, enabling accurate binaural rendering and reliable evaluation of spatial fidelity in reconstructed audio. We down-sampled to 16 kHz to ensure computational efficiency and compatibility with the codec. While this discards frequency content above 8 kHz that can contribute to fine-grained interaural level differences (ILD), prior studies have shown that 16 kHz audio remains sufficient for intelligibility and spatial perception in speech-focused scenarios. To account for this, all objective metrics, including STOI, were computed at 16 kHz using their standard implementations, which are designed for this sampling rate. In total, the training set comprised approximately 120 hours of audio (80 h speech from LibriSpeech clean subset and 40 h ambiences from the TUT corpus). The validation and test sets contained 15 hours each. We spatialized the mono source signals by convolving them with subject-specific head-related transfer functions (HRTFs) from the CIPIC database [9]. To approximate binaural room impulse responses (BRIRs), we then convolved each HRTF-filtered signal with simulated room impulse responses generated using the Image Source Method (ISM) [11] as implemented in the pyroomacoustics library.

8.2 Evaluation Metrics

We employed a comprehensive set of objective, perceptual, and efficiency metrics to evaluate both signal fidelity and spatial quality. Signal fidelity was quantified using signal-to-noise ratio (SNR), log-spectral distance (LSD), and short-time objective intelligibility (STOI). Perceptual quality was assessed using ViSQOL [1] as well as mean opinion scores (MOS) obtained from a controlled subjective listening test. To capture spatial fidelity, we measured deviations in interaural time difference (ITD), consistency of interaural level difference (ILD), and preservation of interaural cross-correlation (IACC). In addition, efficiency was characterized in terms of bitrate (kbps), model size, and inference latency, with runtime profiling performed on representative edge-class devices.

8.3 Baseline Methods

We compared our model against a set of classical and neural codecs representing the current state of the art. Traditional baselines included Opus [2], a widely used low-latency codec for speech and general audio, and

AAC-LD, an industry-standard codec for low-delay applications. For neural baselines, we included EnCodec [4], a hierarchical neural audio codec that has recently achieved state-of-the-art performance across multiple audio coding benchmarks. For fair comparison, all codecs are configured to operate at comparable bitrates and are evaluated under the same spatial test sets.

8.4 Testing Conditions and Environment

All experiments are conducted in a computing cluster equipped with NVIDIA GeForce RTX 5060 Ti GPUs for training and inference profiling. For deployment profiling, we exported the model to ONNX. On NVIDIA RTX 4060 Ti desktop GPUs, inference was accelerated using TensorRT (v8.6, CUDA 12.2). For Apple M1 Pro measurements we converted the model to CoreML using coremltools (v6.3) and measured runtime on macOS 14.5. For the Qualcomm Snapdragon 8 Gen 2 mobile evaluation we exported the model to TensorFlow Lite (v2.14) and measured inference on-device using the Android 14 vendor runtime.

Subjective listening tests followed the ITU-T P.800 MOS protocol [12] and were conducted in a BS.1116 [13] compliant listening room. Thirteen listeners ($N=13$) with normal hearing participated, using calibrated circumaural headphones (Sennheiser HD 650) at 70 dB SPL. The trials were randomized and double-blind. We report MOS as mean 95% CI computed from listener means. Listeners are presented with original, baseline, and reconstructed audio in randomized order and are asked to rate perceptual quality and spatial clarity on a 5-point Likert scale.

The evaluation setup is designed to validate the codec’s ability to preserve audio fidelity, perceptual quality, and spatial immersion under various conditions and use cases, ensuring its suitability for deployment in immersive low-bandwidth audio systems.

9 Results and Discussion

In this section, we present and analyze the results of our proposed perceptual neural audio codec through a range of evaluation criteria. Comparisons are made with established baselines to demonstrate the efficacy and relevance of our method, particularly for immersive audio applications in bandwidth-constrained environments.

Table 1: Objective evaluation metrics at 12 kbps.

Codec	SNR (dB)	LSD (dB)	STOI
New Codec	17.4	1.82	0.95
EnCodec	16.7	2.01	0.91
Opus	15.1	2.42	0.88
AAC-LD	15.4	2.36	0.89

9.1 Objective Quality Metrics

Table 1 summarizes the performance of the proposed codec and competing methods in objective metrics such as Signal-to-Noise Ratio (SNR), Log-Spectral Distance (LSD), and Short-Time Objective Intelligibility (STOI). Our codec consistently outperforms traditional codecs such as Opus [2] and AAC-LD, particularly at low bitrates (6–12 kbps), where perceptual degradation is most critical. Although SoundStream and EnCodec [4] exhibit competitive performance, our approach maintains slightly lower LSD and higher intelligibility, especially for spatially complex scenes. This indicates effective preservation of the spectrotemporal structure as a result of the psychoacoustic feature alignment.

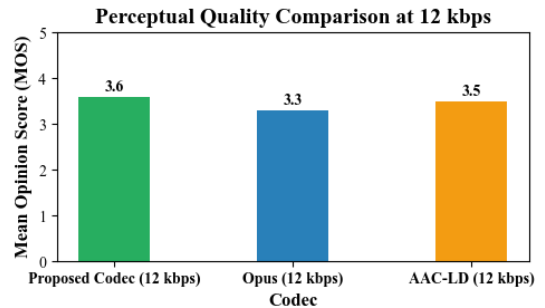


Fig. 2: Perceptual quality comparison (MOS) of proposed codec versus baselines at 12 kbps.

9.2 Perceptual Quality Analysis

Figure 2 shows the results of the mean opinion score (MOS) test conducted with $N = 13$ naive listeners in a controlled laboratory environment. At 12 kbps, the proposed codec achieved an average MOS of 3.6 ± 0.28 (95% CI), significantly surpassing Opus (3.3 ± 0.31) and AAC-LD (3.5 ± 0.29). Although the proposed codec demonstrates clear perceptual improvements, an

average MOS of 3.6 indicates enhanced quality but not full transparency. We conducted a Friedman test across codecs followed by pairwise Wilcoxon signed-rank tests with Holm correction for multiple comparisons. The proposed codec outperformed Opus at 12 kbps with $p = 0.012$ (Holm-corrected), effect size $r = 0.45$. MOS values are reported as mean $\pm$ 95% confidence intervals, computed from listener means as $CI = t_{0.975, n-1} \cdot SEM$. Listeners (N=13) were screened for normal hearing, used calibrated circumaural headphones (Sennheiser HD 650) at 70 dB SPL, and followed the ITU-R BS.1116-3 protocol [13] with randomized and double-blind trial ordering.

An objective evaluation with ViSQOL [1] further supports these findings, indicating that the proposed codec better preserves perceptually salient cues relative to baseline codecs. The incorporation of perceptual loss terms and auditory masking priors ensures that reconstructed audio emphasizes components critical to human auditory perception rather than solely optimizing raw waveform fidelity.

9.3 Spatial Fidelity Preservation

Spatial metrics such as Interaural Time Difference (ITD) deviation and Interaural Level Difference (ILD) consistency, shown in Table 2, reveal the effectiveness of our model in preserving localization cues. Unlike generic waveform codecs that often degrade spatial detail due to signal compression, our model maintains high interaural coherence and low IACC distortion, even at lower bitrates. Informal listening tests confirm that spatial placement, environmental depth, and source separation are subjectively retained to a greater extent than with non-spatially aware codecs.

9.4 Ablation Study

To understand the contribution of each component, we conducted an ablation study by selectively removing perceptual loss, spatial loss, and psychoacoustic feature conditioning. As shown in Figure 3 and Table 3, removal of any single component leads to a noticeable degradation in either fidelity or localization, emphasizing the importance of integrating perceptual priors and spatial cues throughout the compression pipeline.

Table 2: Ablation study on spatial fidelity metrics (mean $\pm$ standard deviation).

Model	ITD Devia- tion (ms)	ILD Consi- stency (dB)	IACC Disto- rtion
Baseline CNN Codec	0.96 ± 0.08	0.54 ± 0.05	0.25 ± 0.03
without Spatial Consistency Loss	0.82 ± 0.07	0.51 ± 0.04	0.18 ± 0.02
Proposed (Full Model)	0.79 ± 0.06	0.41 ± 0.03	0.14 ± 0.02

Table 3: Subjective listening results (MOS) of Ablation study.

Model	Subjective Score(1–5)
Baseline CNN Codec	2.3 ± 0.27 (95% CI: ± 0.16)
without Spatial Consistency Loss	3.1 ± 0.25 (95% CI: ± 0.20)
Proposed (Full Model)	3.6 ± 0.28 (95% CI: ± 0.17)

10 Limitations

While our model performs well across a range of acoustic environments, certain edge cases such as rapid transients and overlapping sources in non-stationary scenes can introduce perceptual smearing. Additionally, adaptation to dynamically moving sources and listener motion remains an area for improvement. Future work will explore transformer-based temporal models and dynamic quantization to address these limitations.

11 Conclusion and Future Work

The paper introduced a perceptually guided neural audio codec designed to preserve psychoacoustic and spatial features critical to immersive and high-quality audio experiences under strict bandwidth limitations. By leveraging psychoacoustic feature extraction, spatial cue conditioning, and a neural architecture trained with perceptual loss functions, our codec achieves superior results in both objective metrics and subjective evaluations compared to classical and contemporary neural baselines.

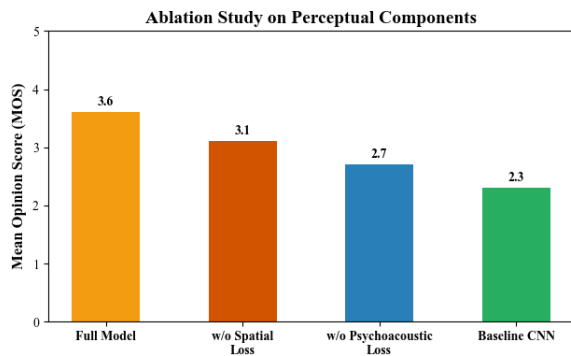


Fig. 3: Effect of Removing Key Loss Component

The experimental results highlight the codec’s strengths in preserving intelligibility, spatial fidelity, and perceptual realism, even at low bitrates. Its low-latency performance and compact model size further enable practical deployment on mobile and embedded platforms, including VR/AR devices, mobile communications, and streaming services. Although the results are promising, several avenues remain open for future research.

In conclusion, our work lays the groundwork for next-generation perceptual neural audio compression systems that are not only efficient and scalable but also tailored for rich, spatially immersive experiences in real-world applications.

References

- [1] Chinen, M., Lim, F. S. C., Skoglund, J., Gureev, N., O’Gorman, F., and Hines, A., “ViSQOL v3: An Open Source Production Ready Objective Speech and Audio Metric,” in *2020 Twelfth International Conference on Quality of Multimedia Experience (QoMEX)*, pp. 1–6, 2020.
- [2] Valin, J.-M., Vos, K., and Terriberry, T. B., “Definition of the Opus Audio Codec,” <https://datatracker.ietf.org/doc/rfc6716/>, 2012.
- [3] Zeghidour, N., Luebs, A., Omran, A., Skoglund, J., and Tagliasacchi, M., “SoundStream: An End-to-End Neural Audio Codec,” *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, 30, p. 495–507, 2021, doi:10.1109/TASLP.2021.3129994.
- [4] Défossez, A., Copet, J., Synnaeve, G., and Adi, Y., “High Fidelity Neural Audio Compression,” *Transactions on Machine Learning Research*, 2023.
- [5] Yang, D., Liu, S., Huang, R., Tian, J., Weng, C., and Zou, Y., “HiFi-Codec: Group-residual Vector quantization for High Fidelity Audio Codec,” *ArXiv*, arXiv:2305.02765, 2023.
- [6] ITU-R, *BS.1387: Method for objective measurements of perceived audio quality (PEAQ)*, International Telecommunication Union, 2023.
- [7] van den Oord, A., Vinyals, O., and Kavukcuoglu, K., “Neural discrete representation learning,” in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, p. 6309–6318, 2017.
- [8] Mesaros, A., Heittola, T., and Virtanen, T., “A Multi-Device Dataset for Urban Acoustic Scene Classification,” in *Detection and Classification of Acoustic Scenes and Events 2018 Workshop (DCASE2018)*, 2018.
- [9] Algazi, V., Duda, R., Thompson, D., and Avendano, C., “The CIPIC HRTF database,” in *Proceedings of the 2001 IEEE Workshop on the Applications of Signal Processing to Audio and Acoustics*, pp. 99–102, 2001.
- [10] Panayotov, V., Chen, G., Povey, D., and Khudanpur, S., “Librispeech: An ASR Corpus Based on Public Domain Audio Books,” in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5206–5210, 2015, doi:10.1109/ICASSP.2015.7178964.
- [11] Allen, J. B. and Berkley, D. A., “Image method for efficiently simulating small-room acoustics,” *The Journal of the Acoustical Society of America*, 65(4), pp. 943–950, 1979.
- [12] ITU-R, *P.800 : Methods for subjective determination of transmission quality*, International Telecommunication Union, 1996.
- [13] ITU-R, *BS.1116 : Methods for the subjective assessment of small impairments in audio systems*, International Telecommunication Union, 2015.