

L3AC: Towards a Lightweight and Lossless Audio Codec

Linwei Zhai¹, Han Ding¹, Cui Zhao¹, Fei Wang¹, Ge Wang¹, Zhi Wang¹, Wei Xi¹

¹Xi'an Jiaotong University, China

Abstract

Neural audio codecs have recently gained traction for their ability to compress high-fidelity audio and provide discrete tokens for generative modeling. However, leading approaches often rely on resource-intensive models and complex multi-quantizer architectures, limiting their practicality in real-world applications. In this work, we introduce L3AC, a lightweight neural audio codec that addresses these challenges by leveraging a single quantizer and a highly efficient architecture. To enhance reconstruction fidelity while minimizing model complexity, L3AC explores streamlined convolutional networks and local Transformer modules, alongside *TConv*—a novel structure designed to capture acoustic variations across multiple temporal scales. Despite its compact design, extensive experiments across diverse datasets demonstrate that L3AC matches or exceeds the reconstruction quality of leading codecs while reducing computational overhead by an order of magnitude. The single-quantizer design further enhances its adaptability for downstream tasks. The source code is publicly available at <https://github.com/zhai-lw/L3AC>.

1 Introduction

Recent advances in neural audio codec technologies have markedly improved the compression and reconstruction of high-fidelity audio. Unlike traditional audio codecs, systems based on deep neural networks not only compress audio efficiently but also produce discrete codes that can be utilized as tokens in sound language modeling (LM) (Wu et al. 2024). This dual functionality underscores their critical importance in modern audio processing tasks. By integrating tokenized outputs of neural audio codecs with language models, a seamless bridge is formed between audio compression and generative language modeling, opening the door to a variety of novel applications.

Despite these benefits, many state-of-the-art (SOTA) neural codecs (Zeghidour et al. 2021; Yang et al. 2023; Du et al. 2024; Défossez et al. 2023; Kumar et al. 2024) rely on complex, multi-quantizer architectures to achieve their high performance in compression and reconstruction. These systems typically employ multiple quantizers arranged hierarchically, each capturing different levels of detail in the audio signal. While this structure enhances reconstruction fidelity, it introduces two significant limitations. First, the resultant hierarchical token streams necessitate customized aggregation techniques to support downstream models (Copet et al. 2023;

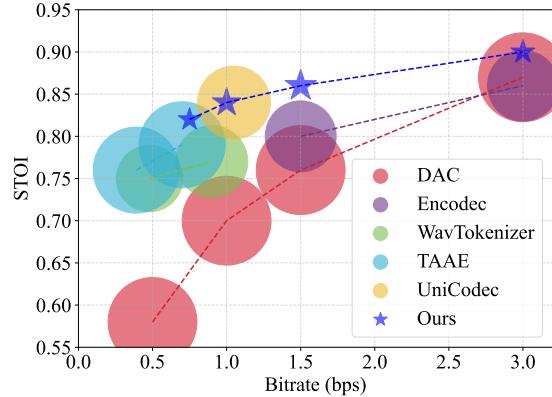


Figure 1: Comparison of various audio codecs. The y-axis (STOI) indicates the quality of the reconstructed audio, where higher values denote better performance. The x-axis (Bitrate) represents the compression bitrate—lower values correspond to smaller compressed audio sizes and higher compression efficiency. Circle size denotes model complexity: smaller sizes represent more lightweight and faster models.

Ji et al. 2025), complicating both training and inference. Second, as the number of quantizers increases, inconsistencies can emerge in the generated tokens. These inconsistencies make it challenging for language models to reliably predict subsequent tokens (Liu et al. 2024).

Although some multi-quantizer frameworks can function as a single quantizer by utilizing only one of their components, this often sacrifices reconstruction quality. In response, recent efforts (Li et al. 2024; Parker et al. 2025; Ji et al. 2025) have explored single-quantizer neural codecs. These codecs offer a more streamlined alternative, yet they show only modest improvements in reconstruction quality over their multi-quantizer counterparts and often rely on resource-intensive architectures, which limits their practical applicability.

To tackle the above challenges, we propose L3AC, a Lightweight and LossLess neural Audio Codec. L3AC employs one quantizer for discrete token generation, eliminating the need for hierarchically structured tokens. Our design leverages lightweight convolutional networks and local Transformer modules to enhance feature extraction and processing efficiency. Furthermore, we introduce *TConv*, a novel module designed to capture both short- and long-term acoustic varia-

tions. This module enables high-fidelity audio reconstruction while maintaining low computational overhead.

The advantages of L3AC are multifaceted. By employing just one quantizer and significantly reducing computational and memory requirements, it can be easily and cost-effectively integrated into diverse audio processing applications. Our experimental results also demonstrate that, despite its architectural simplicity, L3AC outperforms SOTA methods in terms of both audio quality and token generation quality. Comprehensive evaluations across diverse datasets further validate its robustness, adaptability, and efficiency, making it a compelling choice for modern audio tasks.

We summarize the key contributions of this work as follows:

- **Efficient Model Architecture:** The proposed design achieves reduced parameter count and computational cost by an order of magnitude while maintaining high performance.
- **Single-Quantizer Design:** L3AC eliminates the need for hierarchical quantization, simplifying token generation and downstream integration.
- **Enhanced Feature Extraction for Audio Data:** We introduce TConv, a novel module designed to capture both short- and long-term acoustic variations, facilitating higher-fidelity audio reconstruction.

2 Related Works

2.1 Neural Audio Codecs

Traditional audio codecs (Valin, Vos, and Terriberry 2012; Dietz et al. 2015; Neuendorf et al. 2013) mainly rely on signal processing techniques such as linear predictive coding and transform coding for compressing audio signals. While effective, these methods often depend on manually engineered designs, which can limit their flexibility and applicability. The emergence of neural audio codecs has introduced a paradigm shift by adopting data-driven frameworks that learn efficient audio representations from large datasets. These neural models (Zeghidour et al. 2021; Yang et al. 2023; Défossez et al. 2023; Ai et al. 2024; Kumar et al. 2024; Du et al. 2024; Xu et al. 2024; Li et al. 2024) typically employ an encoder-decoder architecture. The encoder compresses input audio into a compact latent feature and then quantizes it into a discrete representation, while the decoder reconstructs the audio from this representation. A significant milestone in this domain was achieved by SoundStream (Zeghidour et al. 2021), which introduced a fully convolutional encoder-decoder network integrated with a residual vector quantizer (RVQ). This innovation enabled the unified handling of diverse audio types, such as speech and music, demonstrating robust performance across various domains.

2.2 High-Fidelity Multi-Quantizer Audio Codecs

Recent advancements in audio codec have increasingly focused on multi-quantizer architectures (Zeghidour et al. 2021; Yang et al. 2023; Du et al. 2024; Défossez et al. 2023; Kumar et al. 2024) to improve reconstruction fidelity and minimize compression errors. EnCodec (Défossez et al. 2023) pushed

the boundaries of performance by employing a sophisticated network architecture and introducing a novel loss design. Building on these efforts, DAC (Kumar et al. 2024) introduced further innovations, including the Snake activation function, the improved RVQ, and a larger network design. In addition, DAC optimized both adversarial and reconstruction loss functions, achieving SOTA performance in audio compression. However, despite their impressive results, these models are computationally intensive and challenging for integration into other types of downstream tasks due to their multi-quantizer architecture, which often limits their applicability in real-world scenarios.

2.3 Single-Quantizer Audio Codecs

Single-quantizer audio codecs have recently garnered significant interest for their ability to produce a single, unified stream of discrete tokens, simplifying integration with downstream generative models. However, many prominent models are optimized for high-fidelity reconstruction within specific domains. For instance, SingleCodec (Li et al. 2024) and TAAE (Parker et al. 2025) demonstrate strong performance on speech, but their generalization to diverse, out-of-domain audio remains a challenge. To address this limitation, approaches like WavTokenizer (Ji et al. 2025) and UniCodec (Jiang et al. 2025) have been developed for general-purpose audio types. However, this broader scope often comes at the cost of reconstruction accuracy, with objective metrics such as PESQ (Rix et al. 2001) and STOI (Taal et al. 2010) falling short. Furthermore, these models tend to be computationally intensive, limiting their practical applicability.

Consequently, a clear research gap exists for a single-quantizer codec that simultaneously achieves high objective fidelity, cross-domain generalizability, and computational efficiency. Our proposed model, L3AC, is designed to fill this void by introducing a lightweight, streamable architecture that excels in high-quality reconstruction across varied audio types.

3 L3AC

L3AC is a lightweight and efficient neural audio codec that delivers high-fidelity audio with low computational complexity and strong scalability. This section outlines the design components of L3AC, beginning with a discussion of its core structure aimed at achieving a sufficient receptive field, followed by a detailed overview of the complete model.

3.1 Efficient Receptive Field Expansion

One of the key challenges in designing effective neural audio codecs lies in achieving a sufficiently large receptive field to capture both fine-grained details and long-range signal dependencies. This capacity is critical for preserving audio quality, as demonstrated by our ablation study.

Conventional approaches such as DAC (Kumar et al. 2024) and EnCodec (Défossez et al. 2023) rely on deep convolutional stacks to enlarge the receptive field, which introduces considerable computational overhead during training and inference. Other solutions based on Recurrent Neural Network (RNN) or transformer architectures (Li et al. 2024; Parker

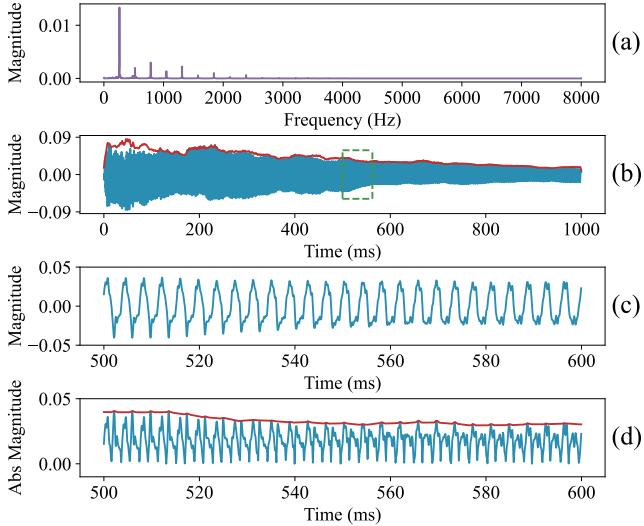


Figure 2: Visualization illustrating the impact of **TPooling** on audio signals, showcasing its effectiveness in processing and capturing long-term signal-level features.

et al. 2025) effectively capture global context but still suffer from limited parallelization or substantial computational demands. To address these limitations, we propose a hybrid architecture that combines convolutional layers with local transformers. This design achieves comparable receptive fields with fewer layers, enabling lower latency and reduced complexity, while remaining suitable for streaming scenarios.

Critically, receptive field expansion should be considered at both the **acoustic** and **signal** levels. As shown in Figure 2, we analyze a one-second audio clip from a piano¹. This audio primarily consists of a fundamental frequency and its overtones, producing distinct time-domain signal variations shown in Figure 2(b). Zooming into a specific segment, highlighted by the green dotted square in Figure 2(c), reveals both short-term variations (at the sampling-point scale) and long-term trends (approximately 10 milliseconds). While the hybrid network allows deeper layers to extract high-level acoustic representations, shallow convolutional layers still struggle to model longer signal-level patterns, e.g., the 10ms trends mentioned above. Dilated convolutions (Holschneider et al. 1990; Shensa 1992; Yu and Koltun 2016) have been explored to address this, but their effectiveness is often diminished by the periodic nature of audio signals, where dominant short-term variations overshadow the long-term trends.

To bridge this gap, we introduce **TPooling**, a novel pooling structure designed to explicitly capture global amplitude variations. Formally, TPooling is defined as:

$$T\text{Pooling}(x, K) = \text{AvgP}(\text{MaxP}(|x|, K), K), \quad (1)$$

where $|x|$ represents the absolute value of the input signal, K denotes the kernel size, and AvgP and MaxP refer to average and maximum pooling operations, respectively.

As shown in Figure 2(d), the blue line represents $|x|$, while the red line shows the result of $T\text{Pooling}(x, K)$. Similarly, in Figure 2(b), the red line illustrates the TPooling output

over the full audio signal. It can be seen that this two-stage pooling process effectively smooths over short-term fluctuations while preserving longer-term patterns; therefore, we propose incorporating TPooling with other network modules to enhance its ability to capture long-term signal dynamics essential for high-quality audio reconstruction.

3.2 Model Design

The architecture of L3AC consists of four main components: an encoder, a quantizer, a decoder, and a discriminator. Each is tailored to balance fidelity, efficiency, and adaptability, ensuring suitability for a wide range of deployment scenarios.

Encoder. The encoder converts raw audio waveforms into compact, informative representations that capture temporal patterns across multiple scales. The encoding process begins with the *TConv Unit*, which is designed to establish a large receptive field at the signal level from the outset. It first applies *TPooling* with varying kernel sizes, followed by a convolutional layer to capture temporal variations at different resolutions. These multi-scale features are concatenated and processed by a point-wise convolution that expands the channel dimension fourfold. The result is passed through a GELU activation function to generate latent representations. These are then concatenated with the original input and compressed back to the original channel size via another point-wise convolution, producing the final TConv output.

Next, the output is passed through a series of *Conv Units* and *Down Layers*. The Conv Unit is an adaptation of the ConvNeXt architecture (Liu et al. 2022) for time-domain audio processing. Here, 2D convolutions are replaced with 1D convolutions. Additionally, the Snake activation function (Ziyin, Hartwig, and Ueda 2020) is employed to effectively capture the periodic and nonlinear characteristics of audio, as demonstrated in prior studies (Kumar et al. 2024; Lee et al. 2022). This design achieves a balance between computational efficiency and high-fidelity feature extraction. Following the Conv Unit, the Down Layer applies a strided convolutional network to perform downsampling, a method commonly adopted in previous studies (Défossez et al. 2023; Kumar et al. 2024; Yang et al. 2023). This reduces the data resolution while preserving critical information, improving processing efficiency in subsequent stages.

Finally, the downsampled features are fed into a *Local Transformer*, which provides a large acoustic-level receptive field with low computational overhead and latency. To maintain causality and ensure real-time performance, the transformer avoids backward dependencies. Its self-attention mechanism dynamically prioritizes relevant segments of the input, yielding high-quality acoustic representations. Notably, for low-bitrate models, an additional downsampling layer is added between the Local Transformer layers to further compress the audio.

Quantizer. At the core of L3AC is a single quantizer used to discretize the features extracted by the encoder. This design choice simplifies the model, reduces resource consumption, and enables seamless integration into downstream tasks. However, single-quantizer architectures inherently produce smaller codebooks compared to multi-quantizers, which can

¹[https://en.wikipedia.org/wiki/C_\(musical_note\)](https://en.wikipedia.org/wiki/C_(musical_note))

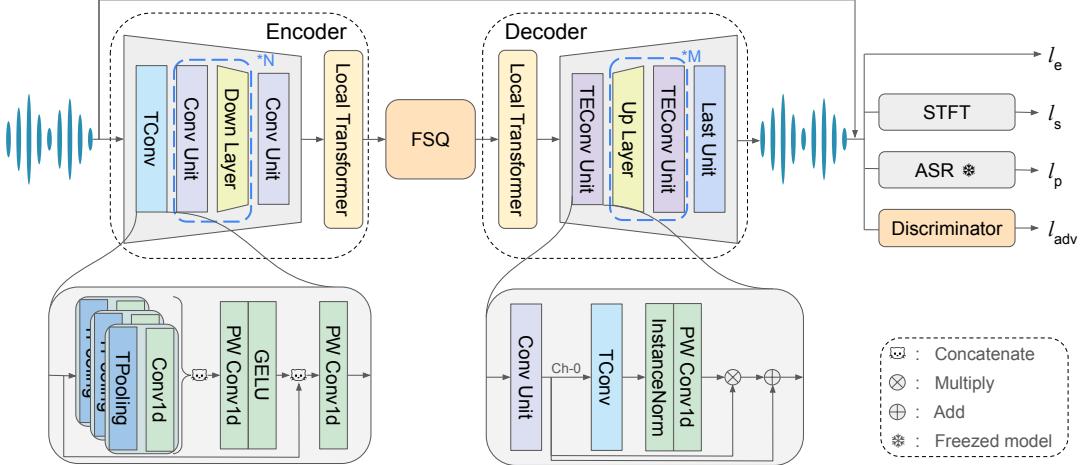


Figure 3: Overview of the L3AC architecture, comprising an encoder, a quantizer (using Finite Scalar Quantization, FSQ), a decoder, and a discriminator. The model is optimized using reconstruction losses (l_e , l_s), perceptual loss (l_p), and adversarial loss (l_{adv}).

significantly degrade model accuracy (Kumar et al. 2024). Additionally, traditional single quantizers often fail when codebooks become large, leading to collapse. To mitigate this, we adopt Finite Scalar Quantization (FSQ) (Mentzer et al. 2023), which supports large codebooks without risking collapse issues. FSQ uses straightforward scalar quantization in a bounded low-dimensional space, reducing complexity, providing faster quantization speeds and greater stability during training.

Additionally, inspired by (Parker et al. 2025), we adopt a hybrid quantization strategy. With a 50% probability, features output by the encoder are perturbed with uniform noise (Brendel et al. 2024) rather than being quantized.

Decoder. The decoder reconstructs high-quality audio from quantized representations. While it mirrors the encoder’s architecture, it incorporates several enhancements to improve reconstruction fidelity.

The decoding process begins with a Local Transformer, which processes the quantized features to restore long-range contextual information. This helps reestablish the acoustic coherence lost during compression. Next, the features are passed through a *TEConv Unit*, which extends the encoder’s Conv Unit by introducing temporal attention. Specifically, it applies a TConv layer along the temporal axis to extract multi-scale latent representations. These are normalized using InstanceNorm and passed through a point-wise convolution to generate temporal attention weights. The original Conv Unit output is then modulated using these weights via an attention-like mechanism, improving the temporal accuracy and perceptual quality of the reconstruction.

Following temporal enhancement, the features undergo resolution restoration through the *Up Layer*. Instead of using transposed convolutions—which are prone to introducing artifacts—this layer adopts linear upsampling to preserve audio integrity while efficiently recovering the original resolution. Finally, the *Last Unit* completes the reconstruction. This com-

ponent consists of deeper convolutional layers with increased parameter capacity and Snake activation functions. The additional parameters allow the model to recover fine-grained details, further enhancing the naturalness and overall quality of the reconstructed audio.

Discriminator. To encourage realistic audio generation, L3AC incorporates the multi-scale discriminator architecture used in DAC (Kumar et al. 2024). This discriminator operates on both time-domain and frequency-domain signals, thereby providing stronger gradient feedback to L3AC’s generator (*i.e.*, the encoder and decoder) and helping mitigate periodicity artifacts, as proved in prior work (Jang et al. 2021; Lee et al. 2022). To ensure balanced training, the discriminator is updated less frequently (*e.g.*, every 15–25 steps), preventing it from converging faster than the generator.

3.3 Training Strategy

L3AC is optimized using a combination of four loss functions (formally defined in Appendix A) to enable lossless audio reconstruction: (1) **Element-Wise Loss** (l_e) evaluates time-domain signal reconstruction by comparing the generated audio with the input audio. (2) **Spectrogram Loss** (l_s) ensures fidelity across multiple frequency scales. (3) **Perception Loss** (l_p) compares intermediate features from an automatic speech recognition (ASR) model (*e.g.*, Whispertiny (Radford et al. 2023)) between original and reconstructed audio, preserving perceptual quality. (4) **Adversarial Loss** (l_{adv}) is calculated by the discriminator, which guarantees realistic audio generation.

The influence of these losses evolves over training. Early stages prioritize l_e and l_s , while later stages emphasize l_p and l_{adv} to refine perceptual quality. However, all four losses tend to decrease simultaneously. To manage the varying impact of each loss while avoiding manual adjustment of weights at different training stages, we employ a loss weight clamping

Model	Bitrate (kbps)	Frame rate (fps)	#Q	MACs↓ (G)	SDR↑	MEL↓	STOI↑	PESQ↑	Audio SDR↑	Audio MEL↓
DAC	0.50	50	1	556.00	-7.93	1.95	0.58	1.08	-10.31	2.57
TAAE†	0.39	25	1	374.89	0.86	1.51	0.76	1.53	-7.86	2.45
TAAE†	0.70	25	2	374.89	2.83	1.45	0.79	<u>1.64</u>	<u>-7.02</u>	2.42
WavTokenizer	0.48	40	1	34.26	-2.80	<u>1.14</u>	0.75	1.51	-11.37	<u>1.44</u>
WavTokenizer†	0.9	75	1	64.17	-1.31	0.97	0.77	1.58	-12.29	1.42
L3AC	0.75	44	1	1.55	1.83	1.14	0.82	1.68	-2.16	1.89
DAC	1.0	50	2	556.01	-4.99	1.31	0.70	1.18	<u>-7.47</u>	2.05
UniCodec	1.05	75	1	71.22	<u>-0.70</u>	0.86	0.84	1.91	<u>-8.24</u>	1.26
L3AC	1.0	59	1	2.03	2.88	1.09	0.84	<u>1.77</u>	<u>-0.99</u>	1.86
Encodec	1.5	75	2	55.95	<u>1.47</u>	1.30	0.80	<u>1.50</u>	<u>-0.58</u>	1.51
DAC	1.5	50	3	556.02	-3.23	<u>1.11</u>	0.76	1.31	-5.36	1.89
L3AC	1.5	89	1	2.37	4.34	1.02	0.86	1.91	0.92	<u>1.83</u>
Encodec	3.0	75	4	55.95	<u>4.42</u>	1.15	0.86	1.91	<u>1.78</u>	1.41
DAC	3.0	50	6	556.05	-0.77	0.83	<u>0.87</u>	<u>1.92</u>	-1.43	1.67
L3AC	3.0	167	1	1.64	6.75	<u>0.92</u>	0.90	2.31	3.89	1.74

Table 1: **Signal-level evaluation results** for various codec models. #Q refers to the number of quantizers used in the model. MACs (Multiply-Accumulate Computations) indicate the total number of arithmetic operations required to process 10 seconds of audio. † denotes that the model is trained specifically for speech.

mechanism (see Equation (2)).

$$clamp(l, max) = \begin{cases} l, & l < max \\ l \times max, & l \geq max \\ l.detach(), & \end{cases} \quad (2)$$

This approach limits the maximum value of certain losses (*e.g.*, l_p , l_{adv}) in the early stages, while allowing other losses (*e.g.*, l_e , l_s) to become more influential, and these effects are reversed once the network enters later stages of training, *i.e.*, when the reconstructed audio quality has improved to a certain extent. Additionally, we adopt the One Cycle Learning Rate policy (Smith and Topin 2019) to dynamically adjust the learning rate, accelerating convergence and improving final performance.

4 Experiments and Results

4.1 Experimental Settings

A summary of our experimental setup is provided below. Additional implementation details can be found in Appendix B.

Training Datasets. L3AC was trained using datasets from three audio domains: 1) **Speech Domain:** The “train-clean-100” and “train-clean-360” subsets of LibriSpeech (Panayotov et al. 2015) for clean speech, alongside the “cv-corpus-18.0” dataset from Common Voice (Mozilla 2024) for noisy speech. 2) **Music Domain:** The low-quality version of MTG-Jamendo dataset (Bogdanov et al. 2019). 3) **General Audio Domain:** The FSD50K dataset (Fonseca et al. 2022), which includes a wide range of audio categories.

Compared to other works, L3AC was trained on a relatively smaller collection of datasets, as detailed in Appendix B. All audio was uniformly resampled to 16,000 Hz, as our empirical results indicated that higher sampling rates did not yield performance gains and substantially increased the computational load. During training, batches were alternated between

these datasets, with samples randomly selected from each dataset. This approach ensured consistent training time and equal weight for each dataset, facilitating balanced learning across different audio domains.

Training Details. We optimized L3AC using the AdamW optimizer, incorporating a one-cycle learning rate schedule (Smith and Topin 2019). The learning rate warmed up from 5×10^{-5} to a peak of 5×10^{-4} before decaying to 5×10^{-6} . Gradient clipping was employed to stabilize training, with maximum norms of 10,000 for the codec and 10 for the discriminator. Additionally, a weight decay of 1×10^{-5} was applied during discriminator training. The training was conducted on a single NVIDIA RTX 4090 GPU, demonstrating the model’s computational efficiency.

Evaluation Details. Evaluation is conducted from two aspects, *i.e.*, assessing the quality of both the reconstructed audio and the discrete tokens generated by the codec.

Reconstructed Audio Evaluation. This evaluation was conducted using the Codec-SUPERB benchmark (Wu et al. 2024), which provides both signal-level and application-level assessments for the reconstructed audio. Specifically, the datasets and evaluation metrics were sourced from the Codec-SUPERB challenge held at SLT 2024².

1) **Signal-Level Evaluations:** We measured Signal-to-Distortion Ratio (SDR) (Raffel et al. 2014), Perceptual Evaluation of Speech Quality (PESQ) (Rix et al. 2001), Short-Time Objective Intelligibility (STOI) (Taal et al. 2010), and Mel Spectrogram Distance (MEL) across 11 datasets covering speech, music, and general audio.

2) **Application-Level Evaluations:** We incorporate four downstream tasks: automatic speech recognition (ASR) on the LibriSpeech dataset, automatic speaker verification (ASV)

²https://github.com/voidful/Codec-SUPERB/tree/SLT_Challenge

Model	Bitrate (kbps)	Frame rate (fps)	#Q	MACs↓ (G)	WER(ASR)↓ (%)	EER(ASV)↓ (%)	Acc(ER)↑ (%)	Acc(AEC)↑ (%)
DAC	0.50	50	1	556.00	78.12	34.54	22.50	13.90
TAAE†	0.39	25	1	374.89	12.12	15.20	57.43	31.70
TAAE†	0.70	25	2	374.89	9.98	12.49	61.88	34.50
WavTokenizer	0.48	40	1	34.26	16.65	12.10	55.07	42.65
WavTokenizer†	0.9	75	1	64.17	13.21	12.31	58.68	41.55
L3AC	0.75	44	1	1.55	5.01	9.52	65.49	65.00
DAC	1.0	50	2	556.01	20.65	18.94	40.28	26.70
UniCodec	1.05	75	1	71.22	6.65	5.14	70.90	58.40
L3AC	1.0	59	1	2.03	4.51	8.03	68.75	69.65
Encodec	1.5	75	2	55.95	11.19	10.72	57.57	62.70
DAC	1.5	50	3	556.02	9.66	10.99	50.35	40.85
L3AC	1.5	89	1	2.37	4.11	6.42	70.42	72.90
Encodec	3.0	75	4	55.95	4.59	4.32	68.12	81.00
DAC	3.0	50	6	556.05	4.26	3.56	70.00	70.40
L3AC	3.0	167	1	1.64	3.41	3.46	71.81	80.90

Table 2: **Application-level evaluation results** for various codec models. **ASR** represents the automatic speech recognition task, **ASV** represents the automatic speaker verification task, **ER** represents the emotion recognition task, and **AEC** represents the audio event classification task.

on the VoxCeleb dataset (Nagrani et al. 2020), emotion recognition (ER) on the RAVDESS dataset (Livingstone and Russo 2019), and audio event classification (AEC) on the ESC-50 dataset (Piczak 2015).

Generated Tokens Evaluation. Inspired by DASB (Mousavi et al. 2024), we evaluated the quality of the generated tokens on two representative downstream tasks: Automatic Speech Recognition (ASR) with tokens generated by codec models and Text-to-Speech (TTS) using tokens to synthesize audio. For ASR, we trained a transcription model on the LibriSpeech dataset using an LSTM-based architecture from SpeechBrain (Ravanelli et al. 2021). For TTS, we trained a synthesis model on the LJSpeech dataset (Ito and Johnson 2017) using a Transformer-based architecture from SpeechBrain. All models were trained under identical settings (including same hyper-parameters, same software versions and same hardware, etc.) to support fair comparison.

Baselines. We compared L3AC against several SOTA models: the multi-quantizer codecs EnCodec (Défossez et al. 2023) and DAC (Kumar et al. 2024), and the single-quantizer models WavTokenizer (Ji et al. 2025), TAAE (Parker et al. 2025), and UniCodec (Jiang et al. 2025). For EnCodec and DAC, performance was evaluated across multiple bitrates by adjusting the number of RVQ levels. For WavTokenizer, the “large-600-24k-4096” and “large-320-24k-4096” pretrained models were used, as they represent the most extensively trained versions of the model. For TAAE and UniCodec, official pretrained checkpoints were used.

4.2 Results

Reconstructed Audio Quality. We evaluated L3AC’s reconstruction performance against SOTA baseline models across various bitrates, with detailed results presented in Table 1 (signal-level metrics) and Table 2 (application-level metrics). The findings show that L3AC consistently matches

or exceeds the performance of SOTA codecs while operating with significantly lower computational complexity. For a more intuitive comparison, Figure 1 plots STOI scores against model bitrate and model complexity, highlighting L3AC’s exceptional efficiency and effectiveness. Since reconstruction quality is directly tied to the compression rate (*i.e.*, bitrate), we partition our analysis into three distinct bitrate ranges to evaluate L3AC’s performance in the following.

Performance below 1 kbps. In the challenging sub-1 kbps range, L3AC demonstrates a decisive advantage. On application-level tasks, L3AC operating at just 0.75 kbps achieves an ASR WER of 5.01% and an AEC accuracy of 65.0%, outperforming all competing models (Table 2). Its superiority is reinforced by leading signal-level metrics, including an STOI of 0.82, a PESQ of 1.68, and an Audio SDR of -2.16 (Table 1). While its Speech/Audio MEL distance is marginally higher than WavTokenizer and its Speech SDR is slightly behind the speech-specialized TAAE model, L3AC delivers the strongest and most versatile overall performance, excelling across a wide range of tasks while operating at a fraction of the computational cost.

Performance at 1 kbps. At 1 kbps, L3AC remains highly competitive with the top-performing models. Although UniCodec achieves a better PESQ score and MEL distance, L3AC achieves a significantly better SDR (2.88 vs. -0.70) and Audio SDR (-0.99 vs. -8.24). Furthermore, while L3AC’s application-level performance is comparable to UniCodec’s, its efficiency is orders of magnitude greater. UniCodec’s computational demand of 71.22G MACs is over 35 times higher than L3AC’s 2.03G MACs, establishing L3AC as a far more practical and efficient solution for achieving high-fidelity results.

Performance above 1 kbps. In this bitrate range, we compare L3AC against multi-quantizer models like Encodec and DAC. The results demonstrate the superiority of L3AC’s

Model	Bitrate (kbps)	Frame rate (fps)	#Q	MACs↓ (G)	WER(ASR-1)↓ (%)	WER(ASR-2)↓ (%)	WER(TTS)↓ (%)	WER(Total)↓ (%)
TAAE†	0.39	25	1	374.89	33.43	53.84	97.57	61.61
TAAE†	0.7	25	2	374.89	27.87	49.16	97.51	58.18
DAC	0.5	50	1	556.00	78.90	89.66	<u>26.93</u>	65.19
DAC	1.0	50	2	556.01	63.17	83.12	35.69	60.66
DAC	1.5	50	3	556.02	57.34	80.29	41.25	59.63
WavTokenizer	0.48	40	1	34.26	64.23	84.23	26.20	58.22
WavTokenizer†	0.9	75	1	64.17	59.51	82.47	36.48	59.49
UniCodec	1.05	75	1	71.22	49.35	75.86	35.03	53.41
L3AC	0.75	44	1	1.55	41.41	68.70	27.83	45.98
L3AC	1.0	59	1	2.03	38.61	66.76	27.18	44.18
L3AC	1.5	89	1	2.37	<u>39.55</u>	<u>68.40</u>	27.98	<u>45.31</u>

Table 3: **Generated Tokens Evaluation** on various codec models. **ASR-1** denotes the ASR task on the LibriSpeech test-clean subset, **ASR-2** is on the test-other subset, **TTS** represents the text-to-speech task, and **WER(Total)** is the mean of the Word Error Rates from all three tasks.

single-quantizer architecture, as it outperforms these more complex models on nearly every metric. At the signal level, L3AC establishes a commanding lead in PESQ scores; for instance, at 3.0 kbps, its PESQ of 2.31 far exceeds Encodec’s 1.91 and DAC’s 1.92. This trend continues in application-level evaluations, where L3AC is significantly better than Encodec and DAC across all settings, with only one trivial exception: its AEC accuracy at 3 kbps (80.9%) is negligibly lower than Encodec’s (81.0%). This proves that L3AC’s design is fundamentally more efficient, capable of reconstructing lossless audio in a single stream of discrete units where other models require multiple, complex layers of quantization.

Across all benchmarks, MEL distance is the only metric where L3AC does not consistently lead. This is an expected outcome of our model’s design. While competing models directly optimize using Mel spectrum-based loss, L3AC employs an STFT-based loss function that prioritizes a broader set of spectral features.

Generated Tokens Quality. Beyond reconstruction, we evaluated the quality of L3AC’s generated tokens for downstream tasks, including automatic speech recognition (ASR) and text-to-speech (TTS), with results shown in Table 3.

It is important to note that the TAAE model was specifically trained on the LibriSpeech dataset, which is used for the ASR-1 and ASR-2 evaluations. Unsurprisingly, it performs well on these tasks but fails catastrophically on the TTS task which uses the out-of-domain LJSpeech dataset ($\text{WER} > 97\%$), revealing its poor generalization.

Excluding the TAAE model, L3AC leads in ASR and remains highly competitive in TTS. At 1.0 kbps, it achieves a TTS WER of 27.18%, narrowly trailing WavTokenizer (26.20%) but outperforming other models, *e.g.*, UniCodec (35.69%), by a significant margin. This well-rounded performance indicates that L3AC’s tokens are semantically rich and generalize effectively across different tasks and datasets.

Our analysis also confirms that increasing the number of quantizers in token-based codecs degrades their usability in downstream applications, *e.g.*, TTS task, reaffirming the architectural limitations of such models. L3AC’s single-

quantizer approach, by contrast, produces a simplified token sequence that enhances compatibility with standard downstream models.

Further analysis reveals a trade-off between frame rate and bitrate. Lower frame rates reduce sequence length, easing the burden on downstream models, but also limit information density due to a lower bitrate if the codebook size is fixed. Our experiments suggest that a bitrate of around 1 kbps with a 60 fps frame rate offers an effective compromise, maximizing utility while maintaining manageable sequence lengths.

Ablation study. Our ablation studies demonstrate the critical role of each design choice. In particular, shrinking the Local Transformer window or replacing TConv with a standard convolution causes severe performance drops or even convergence failures, underscoring the central importance of a sufficiently large receptive field—an aspect seldom emphasized in prior work. Likewise, omitting the perceptual loss term or disabling our loss weight clamp mechanism during training yields a marked degradation across evaluation metrics, validating the necessity of our loss design. Together, these results confirm that both our architectural innovations and training designs are indispensable for achieving robust, high-quality audio modeling. More details can be found in Appendix C.

5 Conclusion

This work introduces L3AC, a lightweight and lossless audio codec, which effectively addresses fundamental challenges related to scalability, adaptability, and computational efficiency. Our comprehensive experimental results demonstrate that L3AC achieves audio quality on par with, and often exceeding, that of more complex SOTA systems—while operating at a fraction of their computational cost. This inherent efficiency, coupled with a simplified training pipeline, substantially reduces the training burden and facilitates rapid adaptation for specialized audio tasks. Consequently, L3AC emerges as a superior choice for researchers and developers seeking an efficient, high-quality, and adaptable neural audio compression solution.

References

- Ai, Y.; Jiang, X.-H.; Lu, Y.-X.; Du, H.-P.; and Ling, Z.-H. 2024. APCodec: A Neural Audio Codec with Parallel Amplitude and Phase Spectrum Encoding and Decoding. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 32: 3256–3269.
- Bogdanov, D.; Won, M.; Tovstogan, P.; Porter, A.; and Serra, X. 2019. The MTG-jamendo Dataset for Automatic Music Tagging. In *Proceedings of the International Conference on Machine Learning (ICML)*.
- Brendel, A.; Pia, N.; Gupta, K.; Behringer, L.; Fuchs, G.; and Multrus, M. 2024. Neural Speech Coding for Real-Time Communications Using Constant Bitrate Scalar Quantization. *IEEE Journal of Selected Topics in Signal Processing*, 1–15.
- Copet, J.; Kreuk, F.; Gat, I.; Remez, T.; Kant, D.; Synnaeve, G.; Adi, Y.; and Défossez, A. 2023. Simple and controllable music generation. In *Proceedings of the 37th International Conference on Neural Information Processing Systems (NeurIPS)*.
- Défossez, A.; Copet, J.; Synnaeve, G.; and Adi, Y. 2023. High Fidelity Neural Audio Compression. *Transactions on Machine Learning Research*.
- Dietz, M.; Multrus, M.; Eksler, V.; Malenovsky, V.; Norvell, E.; Pobloth, H.; Miao, L.; Wang, Z.; Laaksonen, L.; Vasilache, A.; et al. 2015. Overview of the EVS Codec Architecture. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 5698–5702.
- Du, Z.; Zhang, S.; Hu, K.; and Zheng, S. 2024. FunCodec: A Fundamental, Reproducible and Integrable Open-Source Toolkit for Neural Speech Codec. In *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 591–595.
- Fonseca, E.; Favory, X.; Pons, J.; Font, F.; and Serra, X. 2022. FSD50K: An Open Dataset of Human-Labeled Sound Events. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30: 829–852.
- Holschneider, M.; Kronland-Martinet, R.; Morlet, J.; and Tchamitchian, Ph. 1990. A Real-Time Algorithm for Signal Analysis with the Help of the Wavelet Transform. In *Wavelets*, 286–297. Berlin, Heidelberg: Springer. ISBN 978-3-642-75988-8.
- Ito, K.; and Johnson, L. 2017. The LJ Speech Dataset. <https://keithito.com/LJ-Speech-Dataset/>.
- Jang, W.; Lim, D.; Yoon, J.; Kim, B.; and Kim, J. 2021. UnivNet: A Neural Vocoder with Multi-Resolution Spectrogram Discriminators for High-Fidelity Waveform Generation. In *Proceedings of 34th International Conference on Neural Information Processing Systems (Interspeech)*, 2207–2211.
- Ji, S.; Jiang, Z.; Wang, W.; Chen, Y.; Fang, M.; Zuo, J.; Yang, Q.; Cheng, X.; Wang, Z.; Li, R.; Zhang, Z.; Yang, X.; Huang, R.; Jiang, Y.; Chen, Q.; Zheng, S.; Wang, W.; and Zhao, Z. 2025. WavTokenizer: An Efficient Acoustic Discrete Codec Tokenizer for Audio Language Modeling. In *Proceedings of the International Conference on Learning Representations (ICLR)*.
- Jiang, Y.; Chen, Q.; Ji, S.; Xi, Y.; Wang, W.; Zhang, C.; Yue, X.; Zhang, S.; and Li, H. 2025. UniCodec: Unified Audio Codec with Single Domain-Adaptive Codebook. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (ACL)*.
- Kumar, R.; Seetharaman, P.; Luebs, A.; Kumar, I.; and Kumar, K. 2024. High-Fidelity Audio Compression with Improved RVQGAN. In *Proceedings of the 38th Annual Conference on Neural Information Processing Systems (NeurIPS)*, 27980–27993.
- Lee, S.-g.; Ping, W.; Ginsburg, B.; Catanzaro, B.; and Yoon, S. 2022. BigVGAN: A Universal Neural Vocoder with Large-Scale Training. In *Proceedings of the International Conference on Learning Representations (ICLR)*.
- Li, H.; Xue, L.; Guo, H.; Zhu, X.; Lv, Y.; Xie, L.; Chen, Y.; Yin, H.; and Li, Z. 2024. Single-Codec: Single-codebook Speech Codec towards High-Performance Speech Generation. In *Proceedings of 37th International Conference on Neural Information Processing Systems (Interspeech)*, 3390–3394.
- Liu, W.; Guo, Z.; Xu, J.; Lv, Y.; Chu, Y.; Zhao, Z.; and Lin, J. 2024. Analyzing and Mitigating Inconsistency in Discrete Audio Tokens for Neural Codec Language Models. *arXiv*.
- Liu, Z.; Mao, H.; Wu, C.-Y.; Feichtenhofer, C.; Darrell, T.; and Xie, S. 2022. A ConvNet for the 2020s. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 11966–11976.
- Livingstone, S. R.; and Russo, F. A. 2019. RAVDESS Emotional speech audio.
- Mentzer, F.; Minnen, D.; Agustsson, E.; and Tschannen, M. 2023. Finite Scalar Quantization: VQ-VAE Made Simple. In *Proceedings of the International Conference on Learning Representations (ICLR)*.
- Mousavi, P.; Libera, L. D.; Duret, J.; Ploujnikov, A.; Subakan, C.; and Ravanelli, M. 2024. DASB - Discrete Audio and Speech Benchmark. *arXiv*.
- Mozilla. 2024. Mozilla Common Voice. <https://commonvoice.mozilla.org/>.
- Nagrani, A.; Chung, J. S.; Xie, W.; and Zisserman, A. 2020. Voxceleb: Large-scale speaker verification in the wild. *Computer Speech & Language*, 60: 101027.
- Neuendorf, M.; Multrus, M.; Rettelbach, N.; Fuchs, G.; Robilliard, J.; Lecomte, J.; Wilde, S.; Bayer, S.; Disch, S.; Helmrich, C.; et al. 2013. The ISO/MPEG unified speech and audio coding standard—consistent high quality for all content types and at all bit rates. *Journal of the Audio Engineering Society*, 61(12): 956–977.
- Panayotov, V.; Chen, G.; Povey, D.; and Khudanpur, S. 2015. Librispeech: An ASR Corpus Based on Public Domain Audio Books. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 5206–5210.
- Parker, J. D.; Smirnov, A.; Pons, J.; Carr, C. J.; Zukowski, Z.; Evans, Z.; and Liu, X. 2025. Scaling Transformers for Low-Bitrate High-Quality Speech Coding. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

- Piczak, K. J. 2015. ESC: Dataset for Environmental Sound Classification. In *Proceedings of the 23rd ACM International Conference on Multimedia (MM)*, 1015–1018.
- Radford, A.; Kim, J. W.; Xu, T.; Brockman, G.; McLeavey, C.; and Sutskever, I. 2023. Robust Speech Recognition via Large-Scale Weak Supervision. In *Proceedings of the International Conference on Machine Learning (ICML)*, 28492–28518.
- Raffel, C.; McFee, B.; Humphrey, E. J.; Salamon, J.; Nieto, O.; Liang, D.; and Ellis, D. P. W. 2014. Mir_eval: A Tranparent Implementation of Common Mir Metrics. In *Proceedings of the 15th International Society for Music Information Retrieval Conference (ISMIR)*.
- Ravanelli, M.; Parcollet, T.; Plantinga, P.; Rouhe, A.; Cornell, S.; Lugosch, L.; Subakan, C.; Dawalatabad, N.; Heba, A.; Zhong, J.; Chou, J.-C.; Yeh, S.-L.; Fu, S.-W.; Liao, C.-F.; Rastorgueva, E.; Grondin, F.; Aris, W.; Na, H.; Gao, Y.; Mori, R. D.; and Bengio, Y. 2021. SpeechBrain: A General-Purpose Speech Toolkit. *arXiv*.
- Rix, A.; Beerends, J.; Hollier, M.; and Hekstra, A. 2001. Perceptual Evaluation of Speech Quality (PESQ)-a New Method for Speech Quality Assessment of Telephone Networks and Codecs. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.
- Shensa, M. 1992. The Discrete Wavelet Transform: Wedding the a Trous and Mallat Algorithms. *IEEE Transactions on Signal Processing*, 40(10): 2464–2482.
- Smith, L. N.; and Topin, N. 2019. Super-Convergence: Very Fast Training of Neural Networks Using Large Learning Rates. In *Proceedings of Artificial Intelligence and Machine Learning for Multi-Domain Operations Applications (I3A)*, 369–386.
- Taal, C. H.; Hendriks, R. C.; Heusdens, R.; and Jensen, J. 2010. A Short-Time Objective Intelligibility Measure for Time-Frequency Weighted Noisy Speech. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 4214–4217.
- Valin, J.-M.; Vos, K.; and Terriberry, T. B. 2012. Definition of the Opus Audio Codec. Request for Comments RFC 6716, Internet Engineering Task Force.
- Wu, H.; Chung, H.-L.; Lin, Y.-C.; Wu, Y.-K.; Chen, X.; Pai, Y.-C.; Wang, H.-H.; Chang, K.-W.; Liu, A.; and Lee, H.-y. 2024. Codec-SUPERB: An In-Depth Analysis of Sound Codec Models. In *Proceedings of Findings of the Association for Computational Linguistics ACL*, 10330–10348.
- Xu, L.; Wang, J.; Zhang, J.; and Xie, X. 2024. LightCodec: A High Fidelity Neural Audio Codec with Low Computation Complexity. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 586–590.
- Yang, D.; Liu, S.; Huang, R.; Tian, J.; Weng, C.; and Zou, Y. 2023. HiFi-Codec: Group-residual Vector Quantization for High Fidelity Audio Codec. *arXiv*.
- Yu, F.; and Koltun, V. 2016. Multi-Scale Context Aggregation by Dilated Convolutions. *arXiv*.
- Zeghidour, N.; Luebs, A.; Omran, A.; Skoglund, J.; and Tagliasacchi, M. 2021. SoundStream: An End-to-End Neural Audio Codec. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30: 495–507.
- Ziyin, L.; Hartwig, T.; and Ueda, M. 2020. Neural Networks Fail to Learn Periodic Functions and How to Fix It. In *Proceedings of the 34th Annual Conference on Neural Information Processing Systems (NeurIPS)*, 1583–1594.

L3AC: Towards a Lightweight and Lossless Audio Codec

Supplementary material

A A. Loss Functions

Below is a more concise, clarified presentation of each loss definition.

- **Element-Wise Loss (l_e):**

$$l_e = \|x - \hat{x}\|_1 \quad (3)$$

l_e computes the L1 distance between the ground truth audio x and the reconstructed audio \hat{x} . This loss serves as a fundamental measure of waveform reconstruction accuracy, directly penalizing any sample-level deviation in the time-domain signal. By minimizing this distance, the model is encouraged to generate an output that is numerically close to the original, ensuring a baseline level of fidelity.

- **Spectrogram Loss (l_s):**

$$l_s = \frac{1}{|I|} \sum_{i \in I} (\|S_i(x) - S_i(\hat{x})\|_1 + \|\log_{10}(S_i(x)^2) - \log_{10}(S_i(\hat{x})^2)\|_1) \quad (4)$$

l_s operates in the frequency domain, which is more aligned with human auditory perception. It averages L1-based measures across multiple STFT scales, where S_i uses a window size of 2^i and a hop length of $2^i/4$, and $I = (5, 6, 7, 8, 9, 10, 11)$. The first term ensures the magnitude of spectral components is similar, while the second log-magnitude term better reflects the logarithmic nature of human loudness perception. This multi-scale approach enables the model to capture both fine-grained temporal dynamics (with smaller windows) and precise frequency information (with larger windows).

- **Perception Loss (l_p):**

$$l_p = \|ASR(x) - ASR(\hat{x})\|_2 \quad (5)$$

l_p is defined as the L2 distance between the latent features extracted by a pre-trained ASR model (ASR) for x and \hat{x} . This loss guides the reconstruction towards perceptual realism, particularly for speech. By minimizing the distance in this feature space, we ensure the generated audio preserves essential characteristics for intelligibility, such as phonetic content.

- **Adversarial Loss (l_{adv}):**

$$l_{adv} = \left(\sum_{k=1}^K \|1 - D_k(\hat{x})\|_2 \right) + 2 \times \left(\sum_{k=1}^K \sum_{l=1}^L \|D_k^l(x) - D_k^l(\hat{x})\|_1 \right) \quad (6)$$

l_{adv} is based on a multi-scale discriminator, consistent with the DAC. Here, K denotes the number of discriminators, D_k denotes the k -th discriminator. D_k^l can generate the latent features of the l -th layer of D_k . This encourages L3AC to improve the naturalness of the reconstructed audio at multiple levels of abstraction.

Model	Bitrate (kbps)	MACs (G)	#Params (M)	Encoder rates	Decoder rates	Transformer window size	Codebook levels
L3AC	0.75	1.55	11.29	(6, 5, 4, 3)	(5, 4, 3, 2, 3)	600	(7, 7, 7, 7, 7, 7)
L3AC	1.0	2.03	11.27	(6, 5, 3, 3)	(5, 3, 3, 2, 3)	750	(7, 7, 7, 7, 7, 7)
L3AC	1.5	2.37	11.25	(6, 5, 3, 2)	(5, 3, 3, 2, 2)	600	(7, 7, 7, 7, 7, 7)
L3AC	3	1.64	10.31	(6, 4, 4)	(4, 4, 3, 2)	400	(9, 9, 9, 7, 7, 7)
Ablation study	0.75	1.53~1.55	11.29	(6, 5, 4)	(5, 4, 3, 2)	75~300	(7, 7, 7, 7, 7, 7)

Table 4: Overview of the training configurations. **Encoder rates** represent the downsampling factors for each Down Layer in the Encoder, while **Decoder rates** denote the corresponding upsampling factors for each Up Layer in the Decoder. **Transformer window size** denotes the window size used by the Local Transformer. **Codebook levels** refers to the number of levels within the FSQ Codebook.

B B. Training Details

As discussed in Section 4.1, we implement multiple L3AC variants targeting different compression rates (*i.e.*, bitrates). Table 4 summarizes the training configurations used for each variant. These settings jointly determine each model’s complexity and the resulting bitrate. For instance, to achieve the 0.75 kbps bitrate, we assume a standard input sampling rate of 16 kHz. The encoder uses downsampling factors of (6, 5, 4, 3), resulting in a total downsampling factor of $6 \times 5 \times 4 \times 3 = 360$. This produces a frame rate of $16000 \text{ Hz}/360 = 44.44 \text{ frames/s}$. Each frame is then quantized using a Finite Scalar Quantization (FSQ) codebook with

6 levels, each of size 7. The number of bits required per frame is thus $6 \times \log_2(7) \approx 16.84$ bits. The final bitrate is calculated as $44.44 \text{ frames/s} \times 16.84 \text{ bits/frame} \approx 748 \text{ bps}$, or 0.75 kbps.

In addition, Table 5 provides an overview of the datasets used to train the various competing methods. A key observation is the scale of data used by leading models. DAC and UniCodec leverage the most diverse and extensive data sources, followed closely by WavTokenizer. In contrast, our proposed L3AC is trained on the fewest datasets, with the only exception being TAAE, which is trained exclusively on clean speech. Despite this limited training data, L3AC achieves SOTA performance, as shown in the last row of the table. The relative-STOI metric represents the ratio of a baseline model’s STOI to that of L3AC, with values derived from Table 1 across all bitrate variants. For each method, we select its best STOI score and compare it to the corresponding STOI score from the same-bitrate variant of L3AC. The results highlight the architectural efficiency and effectiveness of L3AC.

Dataset	L3AC	Encodec	DAC	WavTokenizer	UniCodec	TAAE
Common Voice	✓	✓	✓	✓	✓	
LibriSpeech (LibriTTS)	✓			✓	✓	✓
Libri-Light					✓	✓
DNS Challenge		✓	✓			
VCTK			✓	✓	✓	
DAPS			✓			
FSD50K	✓	✓				
AudioSet		✓	✓	✓	✓	
MTG-Jamendo	✓	✓	✓	✓	✓	
MUSDB			✓	✓	✓	
relative-STOI	100%	95.6%	96.7%	93.9%	100%	96.3%

Table 5: Datasets used in different models’ training. ✓ indicates that the corresponding dataset was included in the training process for the model. **relative-STOI** refers to the relative performance in STOI compared to L3AC.

Model	Transformer window size	TConv	Perception Loss ↓	STOI↑
L3AC	300	✓	13.61	0.842
L3AC	150	✓	14.52	0.833
L3AC	75	✓	NC	NC
L3AC	300		13.94	0.828
L3AC	150		NC	NC
L3AC	300	✓	13.61	0.842
w/o Clamping Loss	300	✓	11.72	0.815
w/o Perceptual Loss	300	✓	41.35	0.809
L3AC	150	✓	14.52	0.833
w/o Clamping Loss	150	✓	NC	NC
w/o Perceptual Loss	150	✓	59.92	0.769

Table 6: Ablation study of L3AC’s key components. ✓ indicates the model variant that uses our proposed TConv module instead of traditional convolutions. NC represents the model cannot converge. Performance is evaluated using our **Perception Loss** (l_p), where lower values indicate better perceptual similarity, and **STOI**, an objective metric for speech intelligibility (higher is better).

C C. Ablation Study

We conducted a series of ablation experiments to validate our key design choices, with all models trained on the Speech Domain datasets (LibriSpeech and Common Voice). The results, presented in Table 6, are summarized below.

C.1 C.1 Receptive Field Size

A sufficiently large receptive field is crucial for modeling audio, both at the raw signal level and the higher contextual level.

Acoustic-level Receptive Field: The window size of the local attention module determines the acoustic-level receptive field. As shown in Table 6, progressively reducing this window size from our chosen value leads to a steady decline in

performance, particularly harming perceptual metrics. Critically, reducing the window to 75 resulted in training convergence failure, highlighting the need for a large contextual window to capture long-range dependencies in audio.

Signal-level Receptive Field: The TConv layers are vital for expanding the signal-level receptive field, allowing the model to capture long-term audio variations and effectively reconstruct the high-fidelity audio. Replacing TConv with standard convolutions, which restrict this receptive field, leads to a drop in reconstruction quality. Furthermore, the TConv module improves model robustness; without it, the model’s training becomes more brittle and fails to converge with a local attention window of 150—a window size that is stable in our full model.

C.2 C.2 Loss Function Design

Our training objective is carefully designed to achieve high-quality results and ensure training stability.

Perceptual Loss: Unlike many existing models, L3AC combines a perceptual loss l_p with traditional loss functions (such as element-wise loss l_e , spectrogram loss l_s , and adversarial loss l_{adv}). This additional constraint aims to enhance the perceptual quality of the generated audio, prioritizing real-world quality over minimizing mathematical error. Removing this term, as seen in Table 6, significantly degrades the quality of the reconstruction audio.

Loss Clamp Mechanism: During training, we clamp the weights of different loss components to prevent any single term from dominating the gradient updates. Disabling this mechanism destabilized the training process, causing a substantial drop in performance and even preventing the model from fitting the data effectively. This demonstrates that loss clamping is essential for balancing competing objectives and ensuring robust convergence.

In summary, these ablation studies confirm that L3AC’s architectural components and training procedures are not just beneficial but essential for achieving SOTA lossless audio reconstruction.