BSCodec: A Band-Split Neural Codec for High-Quality Universal Audio Reconstruction

Haoran Wang[1,2], Jiatong Shi[1], Jinchuan Tian[1], Bohan Li[2], Kai Yu[2], Shinji Watanabe[1]*
[1]Carnegie Mellon University [2]Shanghai Jiao Tong University

Abstract

Neural audio codecs have recently enabled high-fidelity reconstruction at high compression rates, especially for speech. However, speech and non-speech audio exhibit fundamentally different spectral characteristics: speech energy concentrates in narrow bands around pitch harmonics (80-400 Hz), while non-speech audio requires faithful reproduction across the full spectrum, particularly preserving higher frequencies that define timbre and texture. This poses a challenge—speech-optimized neural codecs suffer degradation on music or sound. Treating the full spectrum holistically is suboptimal: frequency bands have vastly different information density and perceptual importance by content type, yet full-band approaches apply uniform capacity across frequencies without accounting for these acoustic structures. To address this gap, we propose **BSCodec** (Band-Split Codec), a novel neural audio codec architecture that splits the spectral dimension into separate bands and compresses each band independently. Experimental results demonstrate that BSCodec achieves superior reconstruction over baselines across sound and music, while maintaining competitive quality in the speech domain, when trained on the same combined dataset of speech, music and sound. Downstream benchmark tasks further confirm that BSCodec shows strong potential for use in downstream applications.[1]

1 Introduction

Neural audio codecs (NACs) (Mousavi et al., 2025; Guo et al., 2025; Shi et al., 2024b; Défossez et al., 2022; Kumar et al., 2023; Zeghidour et al., 2021) have revolutionized audio compression and achieved high-fidelity reconstruction. These codecs generate discrete tokens that not only enable efficient transmission but also serve as representations for downstream audio understanding tasks (Wang et al., 2023; Tian et al., 2025). To enhance task-specific performance, recent work (Zhang et al., 2023; Ye et al., 2025) has incorporated external semantic information into codec design, demonstrating improved results on speech-related benchmarks. Additionally, single-codebook approaches (Ji et al., 2024; Xin et al., 2024; Jiang et al., 2025) have achieved excellent reconstruction quality and downstream performance on clean speech through carefully designed codebook structures, encoder-decoder architectures, and other components.

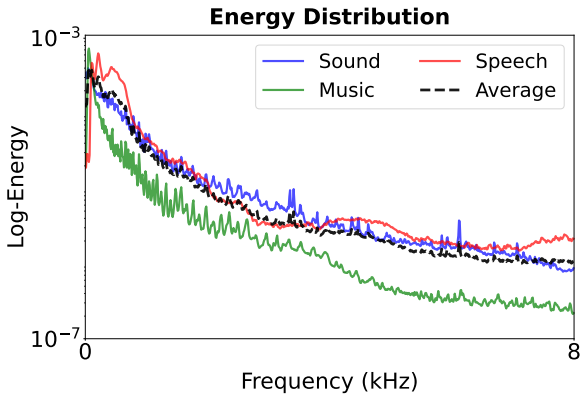

Figure 1: Frequency domain energy distribution of speech, sound and music (extraction method in Appendix D). The distributions exhibit significant structural differences across these three domains, demonstrating distinct spectral characteristics.

However, the real acoustic world would contain components beyond speech, such as sound and music, all with vastly different acoustic characteristics. As illustrated in Figure 1, speech, music, and sound exhibit substantially different average energy distributions across the frequency spectrum. While the aforementioned task-oriented codecs excel on clean speech, their performance degrades significantly (Mousavi et al., 2025) when applied to music and sound domains. Given these spec-

*Corresponding author.
[1]https://github.com/whr-a/espnet/tree/bscodec

Findings of the Association for Computational Linguistics: EACL 2026, pages 4685–4697

tral differences shown in Figure 1, it is reasonable that speech codec designs struggle to generalize to music and sound. Current RVQ-based generalized codecs like DAC (Kumar et al., 2023) still exhibit advantages (Mousavi et al., 2025) in reconstruction quality for universal audio.

While RVQ-based codecs demonstrate reasonable multi-domain performance, their design could be enhanced. The residual quantization hierarchy lacks explicit acoustic structure, with each layer quantizing residuals that do not correspond to interpretable acoustic attributes. This poses challenges for multi-domain compression: as different audio types have distinct energy distributions (Figure 1), a codec handling multiple domains faces a signal source with substantially higher entropy, demanding increased bitrate. Also RVQ applies uniform residual encoding across all domains, requiring the encoder to learn entangled representations for diverse content. Alternative quantization schemes like FSQ (Mentzer et al., 2023) and GroupVQ (Yang et al., 2023) partition along dimensional axes, but these divisions similarly lack grounding in the physical properties of audio signals.

To effectively handle diverse audio domains, we need to account for their structural differences. As shown in Figure 1, different domains exhibit substantially different energy distribution across frequency bands. Therefore, band-splitting offers a natural way to decouple these domain-specific structures by processing each frequency band independently. This approach has proven effective in other audio tasks (Luo and Yu, 2023; Yang et al., 2021), and some RVQ-based multi-band codecs have been successfully developed for specific domains (Ng et al., 2025; Luo et al., 2024), which motivates us to adopt similar decomposition principles for universal audio codec design.

Motivated by these insights, we propose a band-split codec that decomposes input audio into separate frequency bands, each of which is processed independently and in parallel through dedicated encoder, quantizer, and decoder modules. While band-splitting provides a physically grounded design philosophy, its effective application to codec design requires careful consideration of band configuration and quantization allocation. We conduct extensive experiments to systematically investigate the impact of band count, frequency boundaries, and codebook design on multi-domain codec performance. Through principled design of band partitioning and quantization structure, our approach achieves strong performance across all three domains. Specifically, under the same training conditions, our 2.55 kbps and 3.83 kbps models achieve comparable overall performance to DAC at 4.5 kbps and 6 kbps respectively.

Our contributions can be summarized as follows:

- We propose BSCodec, a band-split codec that independently processes time-domain signals from separated frequency bands through parallel encoder-quantizer-decoder modules, achieving strong multi-domain performance through carefully designed band configuration and quantization allocation.
- Comprehensive experiments on reconstruction quality and downstream tasks demonstrate that BSCodec achieves both perceptual superiority and enhanced effectiveness for audio understanding.

2 Related Work

Neural Audio Codecs. The advent of deep learning has enabled significant advances in neural audio compression. They typically adopt the core architecture of RVQ-based discretization paired with GAN (Goodfellow et al., 2014)-driven reconstruction, and further improve performance via targeted architectural enhancements or optimized training strategies (Zeghidour et al., 2021; Défossez et al., 2022; Kumar et al., 2023).

Using a unified model remains challenging due to the distinct energy distribution of universal audio. Recent efforts have pursued universal codecs capable of handling diverse audio types. However, achieving strong multi-domain performance remains challenging due to the distinct acoustic characteristics across domains—speech exhibits narrow-band harmonic structure, music requires wide-band spectral richness, and sound encompasses diverse sound textures (Scharf, 1970). Existing multi-domain codecs address this heterogeneity through various specialized mechanisms. DAC (Kumar et al., 2023) employs balanced sampling strategies, ensuring each training batch contains data from all three domains (speech, music, sound) to prevent domain bias. Further works (Jiang et al., 2025; Liu et al., 2024; Yang et al., 2025) introduce domain-specific codebooks via Mixture-of-Experts layers (Shazeer et al., 2017), semantic priors and MAE-derived representations (He et al., 2022; Huang et al., 2022), and

more fine-grained supervision to further enhance the language modeling capability of universal audio codecs.

In contrast, our work adopts a simpler design philosophy. We employ an architecture and training framework closely aligned with DAC (Kumar et al., 2023), yet achieve comparable performance for both reconstruction and downstream tasks at half the bitrate on the same dataset. This demonstrates that band-split decomposition offers a simple yet effective alternative for multi-domain codec design, obviating the need for complex domain-specific mechanisms while maintaining strong generalization across audio types.

Band-Splitting for Audio Processing. Band-splitting decomposes audio signals into separate frequency bands for independent processing, leveraging the observation that different spectral regions carry distinct perceptual and structural information: low frequencies encode harmonic structure and fundamental pitch, mid frequencies capture formants and timbral characteristics, while high frequencies convey transients and texture. This principle has deep roots in classical signal processing. Subband decomposition techniques such as QMF (Malvar, 1990) and polyphase filter banks (Saramaki and Bregovic, 2002), combined with time-frequency transforms like MDCT (Prokop, 2003), form the foundation for perceptual audio coding standards like MP3 (Brandenburg, 1999) and AAC. Wavelet transforms (Zhang, 2019) similarly provide multi-resolution frequency decomposition.

Recent deep learning approaches have demonstrated the effectiveness of band-split architectures across various audio tasks. In music source separation, Band-split RNN (Luo and Yu, 2023) partitions the spectrogram into multiple frequency bands, processing each band with dedicated recurrent networks before reconstruction. This frequency-aligned decomposition enables the model to specialize in acoustically coherent spectral regions, achieving superior separation quality compared to full-band processing. In neural audio synthesis, Multi-band MelGAN (Yang et al., 2021) employs separate generator branches for different frequency bands, significantly improving synthesis quality and efficiency by exploiting the distinct acoustic properties of each spectral region.

These successes demonstrate that architectures aligned with the natural spectral structure of audio yield improved performance, providing insights for the design of BSCodec.

3 Methodology

The proposed BSCodec follows an encoder-quantizer-decoder architecture with adversarial training, adopting a framework similar to DAC (Kumar et al., 2023). The overall architecture is illustrated in Figure 2.

3.1 Band Splitting

The model first decomposes the input audio waveform into several band-limited signals through time-frequency domain processing. The input discrete-time signal $x[n]$ is transformed into a spectrogram $X(m,k)$ using the Short-Time Fourier Transform (STFT):

$$X(m,k) = \sum_{n=0}^{N-1} x[n+mR]w[n]e^{-j2\pi kn/N} \quad (1)$$

where N denotes the window length (FFT size), n is the sample index within the window, $w[n]$ denotes the window function, R the hop size, m the frame index, and k the frequency bin index.

The band-splitting operation is then performed directly on $X(m,k)$. For an input sampling rate $f_s = 24$ kHz, we define B non-overlapping frequency bands with boundaries $\{f_0, f_1, \ldots, f_B\}$. Each band b is isolated using a binary mask $M_b(k)$:

$$M_b(k) = \begin{cases} 1 & \text{if } f_{b-1} \leq \frac{k \cdot f_s}{N} < f_b \\ 0 & \text{otherwise} \end{cases} \quad (2)$$

The band-specific spectrogram $X_b(m,k)$ is obtained via element-wise multiplication: $X_b(m,k) = X(m,k) \odot M_b(k)$. Each masked spectrogram is then transformed back to the time domain via Inverse STFT:

$$x_b[n] = \sum_m w[n-mR] \times \left(\frac{1}{N} \sum_{k=0}^{N-1} X_b(m,k) e^{j2\pi kn/N} \right) \quad (3)$$

3.2 Encoder and Decoder

The resulting band-limited waveforms $\{x_1[n], x_2[n], \ldots, x_B[n]\}$ are processed in parallel by independent encoders. Each encoder adopts the SEANet (Tagliasacchi et al., 2020) architecture with downsampling strides of $[2, 4, 5, 8]$ across four stages. Starting from an initial channel dimension of 32 that doubles at each stage, the encoders produce latent features with 512 dimensions at a frame rate of 75 Hz. Each convolutional block contains 3 residual units followed by strided convolution. Critically, no parameters are shared

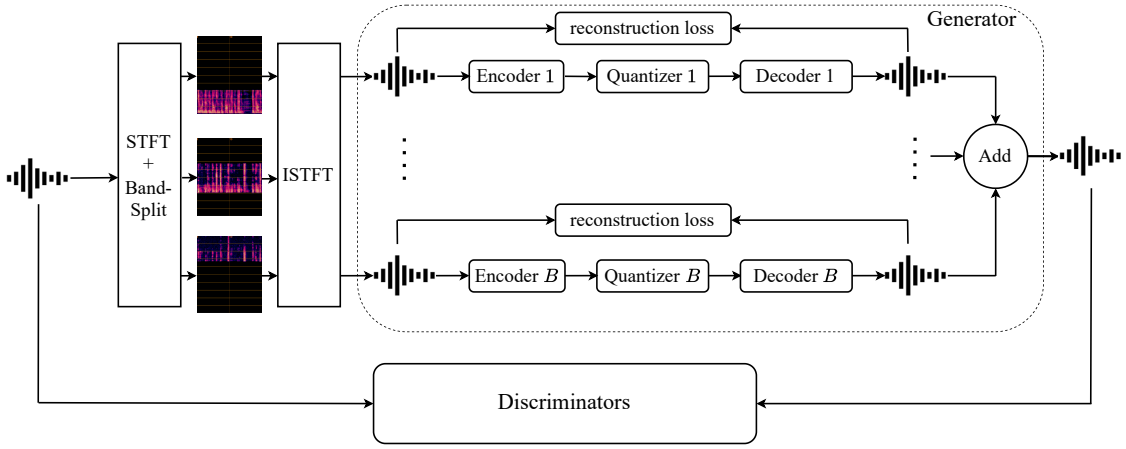

Figure 2: BSCodec architecture with band split, multi-band parallel generators and discriminators.

across the B encoders, allowing each to specialize in the spectral characteristics of its frequency range.

The decoder architecture mirrors the encoder with symmetric upsampling strides of $[8, 5, 4, 2]$, progressively upsampling the quantized representations from 512 dimensions back to 24 kHz waveforms. The final output $\hat{x}[n]$ is obtained by summing all band-specific reconstructions: $\hat{x}[n] = \sum_{b=1}^{B} \hat{x}_b[n]$.

3.3 Vecter Quantization

The vector quantization stage discretizes the continuous latent representations from each band's encoder into discrete tokens. We employ a single-layer quantization scheme per band, utilizing SimVQ(Zhu et al., 2024) to accommodate configurations requiring large codebook capacities.

SimVQ(Zhu et al., 2024) enhances the standard Vector Quantization (VQ) framework by introducing a learnable linear transformation on the codebook embeddings. Given the encoder output z and a codebook of size K: $\{\mathbf{c}_1, \mathbf{c}_2, \ldots, \mathbf{c}_K\}$, the quantization proceeds as follows. First, the nearest codebook entry is determined by:

$$q = \arg \min_{\mathbf{c} \in \{\mathbf{c}_1, \ldots, \mathbf{c}_K\}} \|z - \mathbf{c}W\| \tag{4}$$

where W is a learnable transformation matrix applied to the codebook embeddings. The quantized representation is then computed as:

$$z_q = z + \text{sg}[qW - z] \tag{5}$$

where $\text{sg}[\cdot]$ denotes the stop-gradient operation. During forward propagation, this formulation passes the transformed codebook entry qW to the decoder, while during backpropagation, gradients flow directly through z via the straight-through estimator.

The codebook is optimized through a bidirectional commitment loss with a hyperparameter λ:

$$\mathcal{L}_{\text{commit}} = \|\text{sg}[z] - qW\|^2 + \lambda \|z - \text{sg}[qW]\|^2 \tag{6}$$

The first term encourages the quantized codes to stay close to the encoder outputs, while the second term allows the encoder to adapt towards the codebook. The transformation matrix W improves optimization dynamics by providing a reparameterization that facilitates gradient-based learning of large codebooks. In our experiments, each band is assigned a SimVQ codebook of size $K = 131{,}072$.

3.4 Loss Design

Reconstruction Loss. We adopt a multi-scale mel-spectrogram L1 loss for frequency-domain reconstruction. Specifically, we compute mel-spectrograms at multiple scales ranging from 2^6 to 2^{11} using a Hann window with 80 mel-frequency bins. To provide better supervision signals for each band, we additionally compute the mel loss between each individual band's reconstructed audio and its corresponding ground-truth band, then average these losses and incorporate them into the overall optimization objective.

Adversarial Loss. Following DAC (Kumar et al., 2023), we employ a multi-period discriminator for waveform-level discrimination and a multi-band multi-scale STFT discriminator for frequency-domain discrimination, using the hinge loss formulation with feature matching loss.

Codebook Learning. The codebook commitment loss follows Equation (6) with λ set to 0.25. Gra-

dients are backpropagated through the codebook lookup using the straight-through estimator.

Loss Weighting. We set the loss weighting coefficients as follows: 45.0 for the multi-scale mel loss, 2.0 for the feature matching loss, 1.0 for both the adversarial loss and reconstruction loss, and 1.0 for the codebook commitment loss.

4 Experiments

4.1 Experimental Stages

We conduct our experiments in two progressive stages to systematically validate the effectiveness of band split across different domains.

Stage 1: Music Domain Validation. We first validate the feasibility of band split on music domain. Music is chosen as the initial testbed because its wider fundamental frequency distribution makes it a less challenging learning task for the model. At this stage, we explore various configurations of band split. The experimental results demonstrate that band split achieves excellent performance on music, establishing a strong foundation for further investigation.

Stage 2: Multi-Domain Extension. Building on the success in music, we extend our investigation to a multi-domain setting encompassing speech, music and sound. We aim to determine whether band split maintains its effectiveness when dealing with diverse acoustic characteristics across domains. Our experiments show that band split remains effective in the multi-domain scenario. Furthermore, we introduce targeted optimizations to enhance model performance across all domains.

4.2 Datasets

In Stage 1, we train on Jamendo(Bogdanov et al., 2019), MUSDB18(Rafii et al., 2017), and MAESTRO(Hawthorne et al., 2019). We evaluate on two test sets: the MUSDB18 test set for vocal music, and 100 clips from MAESTRO for non-vocal music.

In Stage 2, we train on approximately 2,100 hours of data spanning three domains. For speech, we use LibriTTS (Zen et al., 2019), VCTK (Yamagishi et al., 2019), and CommonVoice (Ardila et al., 2019) as training data, evaluating on the LibriTTS test-clean. For music, we train on Jamendo(Bogdanov et al., 2019) and MUSDB18(Rafii et al., 2017), evaluating on the MUSDB18 test set. For sound, we train on AudioSet(Gemmeke et al., 2017) and evaluate on the AudioSet test set. To ensure domain balance, We sample about 700 hours from each domain for training. The detailed dataset distribution after sampling is shown in the Appendix E.

4.3 Training Setup

Our model is implemented using the codec part (Shi et al., 2024b) in ESPnet (Watanabe et al., 2018) toolkit. We train with a global batch size of 72, processing audio into 1-second chunks at a sampling rate of 24 kHz. We employ the AdamW (Loshchilov and Hutter, 2017) optimizer with an initial learning rate of 2.0×10^{-4} and momentum coefficients $(\beta_1, \beta_2) = (0.5, 0.9)$. The learning rate is decayed exponentially with a factor of $\gamma = 0.999875$ per epoch, where each epoch processes 2000 samples (approximately 1200 iterations). Training proceeds for 340k iterations in total. For comparison, we train a DAC baseline with 8 codebook layers using identical training configurations and datasets. We use the resulting 6 kbps model, along with its 4.5 kbps and 3 kbps variants obtained via codebook dropout, as baselines for comparison.

4.4 Band Partitioning Configurations

We investigate band partitioning strategies with 5, 3, and 2 frequency bands. The 5-band configuration partitions the spectrum into $[0, 0.5]$, $[0.5, 2]$, $[2, 4]$, $[4, 8]$, and $[8, 12]$ kHz. For coarser granularities, we evaluate a 3-band configuration ($[0, 2]$, $[2, 4]$, $[4, 12]$ kHz) and a 2-band configuration ($[0, 2]$, $[2, 12]$ kHz). Our partitioning strategy is inspired by the frequency division in band-split RNN (Luo and Yu, 2023). While Band-Split RNN uses 41 bands, we adopt only the density distribution of its partitioning. As configurations with fewer bands necessitate larger codebook sizes per sub-band to maintain overall model capacity, we adopt SimVQ(Zhu et al., 2024) for quantization, which efficiently handles large codebooks.

4.5 Evaluation Metrics

We evaluate our model using domain-specific metrics. For speech, we use Mel Cepstral Distortion (MCD), WB-PESQ (Rix et al., 2001), STOI (Taal et al., 2010), Speaker Similarity (SPK_SIM)(Jung et al., 2024) and UTMOS (Saeki et al., 2022). For audio and music, we use VISQOL(Hines et al., 2015), Mel Distance and STFT Distance. All metrics except Mel Distance and STFT Distance are

Table 1: Reconstruction performance of BSCodec and baseline DAC on vocal songs and instrumental music.

Model			Vocal songs			Instrumental music		
Codec	VQ Type	Bitrate	VISQOL↑	Mel Dist.↓	STFT Dist.↓	VISQOL↑	Mel Dist.↓	STFT Dist.↓
DAC	RVQ	6.00 kbps	4.097	0.481	1.018	4.479	0.506	0.959
DAC	RVQ	4.50 kbps	4.075	0.493	1.025	4.459	0.517	0.962
DAC	RVQ	3.00 kbps	4.030	0.515	1.039	4.423	0.541	0.968
BSCodec	5-band VQ	3.75 kbps	**4.384**	**0.445**	**0.921**	4.470	0.419	0.788
BSCodec	3-band SimVQ	3.83 kbps	4.196	0.477	1.042	**4.495**	**0.412**	**0.787**
BSCodec	2-band SimVQ	2.55 kbps	4.104	0.482	0.994	4.484	0.417	0.806

evaluated using the VERSA (Shi et al., 2024a) toolkit with default configurations.

Mel Distance is computed as the L1 distance between mel-scaled magnitude spectrograms using multi-resolution STFT with Hann window lengths ranging from 2^6 to 2^{11} samples projected onto an 80-bin Mel filterbank.

STFT Distance is computed as the L1 distance using multi-scale STFT with a 2048-sample window and 512-sample hop as well as a 512-sample window and 128-sample hop.

5 Results and Discussions

5.1 Music Reconstruction

We first evaluate on the music domain, which exhibits a broad fundamental frequency distribution. We compare against the DAC baselines described in Section 4.3. We assess reconstruction quality on vocal music and solo piano datasets using VISQOL, Mel Distance, and STFT Distance metrics.

As shown in Table 1, our 5-band VQ configuration achieves the best performance on music at 3.75 kbps compared to both DAC baselines, demonstrating that fine-grained frequency decomposition effectively captures harmonic content. On the instrument domain, the 3-band SimVQ achieves competitive results at 3.83 kbps, while the 2-band configuration maintains strong performance at only 2.55 kbps. These results demonstrate that band partitioning is highly effective for domains with wide frequency distributions, with optimal granularity varying by domain complexity.

5.2 Multi-domain Reconstruction

We conduct experiments on a strictly balanced dataset comprising three domains: speech, audio, and music. We train our model alongside a DAC baseline using identical configurations. For speech evaluation, we employ MCD, WB-PESQ, STOI, SPK_SIM and UTMOS as metrics. For audio and music domains, we utilize VISQOL, Mel Distance and STFT Distance.

As shown in Table 2, the speech domain, characterized by a narrow and concentrated fundamental frequency range, poses significant challenges for band splitting. While the 5-band VQ configuration achieves excellent results on audio and music domains, it yields suboptimal performance on speech. Although MCD remains competitive, perceptually-oriented metrics reveal substantial limitations compared to the DAC baseline.

To address this challenge, we adopt a coarser partitioning strategy for speech. Rather than forcing fine-grained low-frequency decomposition, we employ the 3-band SimVQ configuration at 3.83 kbps, which uses a larger codebook to handle a unified low-frequency region while maintaining high-frequency splitting. This modification achieves substantial improvements on speech across all perceptual metrics, with speaker similarity significantly outperforming the DAC baseline. The enhanced speaker similarity demonstrates that our approach effectively preserves speaker-specific characteristics, particularly through the improved low-frequency representation combined with accurate high-frequency reconstruction of formants, fricatives, and plosives.

Notably, the 3-band configuration maintains competitive performance on audio and music domains, with virtually no degradation compared to the 5-band variant. This demonstrates that our adaptive partitioning strategy successfully balances the trade-off between speech-specific requirements and multi-domain effectiveness. The 2-band configuration further reduces bitrate to 2.55 kbps while maintaining reasonable performance across all domains, offering an efficient alternative for bandwidth-constrained scenarios.

5.3 Codebook Utilization

We evaluate codebook utilization on models trained across three domains. We collect statistics on codebook usage frequencies using a test set comprising LibriTTS test-clean and test-other, AudioSet test set, and MUSDB18 test set, totaling 237k seconds

Table 2: Reconstruction performance on speech, sound and music domains. † means the official release.

Model			Speech					Sound			Music		
Codec	VQ Method	Bitrate	MCD↓	PESQ↑	STOI↑	SPK_SIM↑	UTMOS↑	VISQOL↑	Mel Dist.↓	STFT Dist.↓	VISQOL↑	Mel Dist.↓	STFT Dist.↓
EnCodec†	RVQ	6.00 kbps	5.94	2.715	0.939	0.865	3.038	4.240	0.485	0.940	4.410	0.435	0.980
DAC	RVQ	6.00 kbps	5.40	2.915	0.934	0.751	3.356	4.085	0.452	0.874	4.201	0.439	0.974
DAC	RVQ	4.50 kbps	5.50	**2.726**	**0.925**	0.734	3.201	4.055	0.463	0.880	4.171	**0.449**	0.979
DAC	RVQ	3.00 kbps	5.74	2.397	0.905	0.686	2.869	3.990	0.485	0.893	4.105	0.472	0.993
EnCodec†	RVQ	3.00 kbps	6.49	2.048	0.901	0.771	2.305	4.085	0.531	0.978	4.262	0.481	1.014
BSCodec	5# VQ	3.75 kbps	5.08	1.961	0.894	0.810	2.515	**4.245**	0.463	0.800	**4.326**	0.464	0.892
BSCodec	3# SimVQ	3.83 kbps	**5.05**	2.544	0.920	**0.852**	**3.360**	4.234	**0.456**	**0.794**	4.298	0.461	**0.888**
BSCodec	2# SimVQ	2.55 kbps	5.42	2.429	0.916	0.783	3.304	4.137	0.470	0.846	4.166	0.479	0.916

Table 3: Performance comparison on ARCH.

Model		Speech		Audio			Music	
Model	# Enc.	RAVDESS↑	AM↑	ESC-50↑	US8K↑	VIVAE↑	MTT↑	MS-DB↑
DAC	1	0.3958	0.7791	0.3335	0.5311	0.3285	0.2949	0.5754
BSCodec	2	0.4201	0.7801	0.3795	0.5798	0.3265	**0.3555**	**0.7021**
BSCodec	3	0.4306	0.7759	0.3725	0.5654	0.3258	0.3531	0.6969
BSCodec	5	**0.5069**	**0.8548**	**0.3930**	**0.5956**	**0.3810**	0.2580	0.5751

of audio data. As shown in Figure 3, when examining per-layer utilization, our model exhibits slightly lower rates than DAC (98.63% for DAC vs. 92.84% for BSCodec-3band). To analyze inter-codebook correlations, we compute joint utilization by combining codewords from two adjacent codebooks into a larger codebook. The calculation methods for single-layer and joint codebook utilization are detailed in Appendix B. The joint utilization for each pair of adjacent codebooks is shown in Figure 3. The layer-wise growth pattern remains consistent with single-layer observations, but notably, DAC's first two layers exhibit strong correlation, achieving the same joint utilization to our single-layer SimVQ. Moreover, DAC's final two layers show very high joint utilization. However, reconstruction results demonstrate that DAC's final two layers, despite exhibiting very high joint utilization (totaling 20 bits), contribute minimally to performance improvement. In contrast, when we reduce our model from 3-band to 2-band by removing the 17-bit codebook in the high-frequency region, speech performance degrades only slightly at 2.55 kbps, but notable performance gaps emerge in both audio and music domains compared to the 3-band configuration. This suggests that our band-specific codebooks capture domain-critical information more effectively than residual quantization approaches.

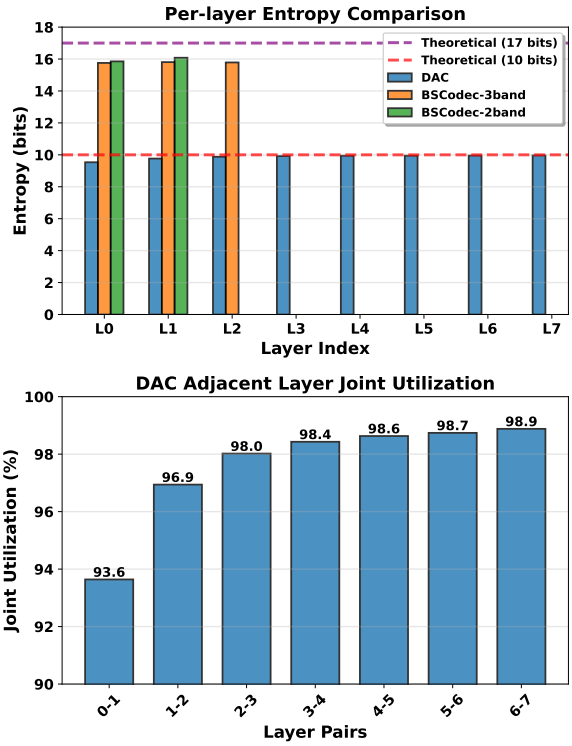

Figure 3: Codebook utilization analysis. Top: Per-layer entropy comparison across different models. Bottom: Joint utilization of adjacent layers in DAC.

5.4 Downstream task

Codec-SUPERB We evaluate our model on downstream tasks from the Codec-SUPERB benchmark (SLT Challenge version (Wu et al., 2024)). We assess performance on three speech tasks: Emotion Recognition (ER, measured by accuracy), Automatic Speaker Verification (ASV, measured by EER), and Automatic Speech Recognition (ASR, measured by WER), as well as Audio Event Classification (AEC, measured by mAP) for the audio domain.

As shown in Table 4, our 3-band BSCodec achieves the best performance. It demonstrates exceptional results on ASV, substantially outperforming both DAC variants, aligning with superior speaker similarity scores. The model excels on AEC and ER, significantly surpassing all base-

lines, while maintaining competitive ASR performance. Notably, the 5-band configuration shows lower performance on speech tasks, suggesting overly fine-grained partitioning fragments speech-relevant information. This confirms 3-band partitioning strikes optimal balance for multi-domain codec design.

Table 4: Performance on Codec-SUPERB. The number before # denotes the number of bands.

Model	Bitrate	Speech			Audio
		ACC↑ (ER)	EER↓ (ASV)	WER↓ (ASR)	mAP↑ (AEC)
DAC	6.00	72.36	3.94	**4.10**	78.15
DAC	4.50	71.74	4.46	4.30	74.90
BSCodec 5#	3.75	67.01	4.01	4.73	82.85
BSCodec 2#	2.55	71.11	4.18	4.62	85.00
BSCodec 3#	3.83	**73.26**	**2.67**	4.40	**87.95**

ARCH We evaluate our model on the ARCH(La Quatra et al., 2024) benchmark, which assesses the semantic richness of codec representations across multiple domains. Following the standard ARCH protocol, we extract and freeze the encoder of each model, append a single linear classification layer, and train for 1000 epochs until full convergence. The speech domain includes the RAVDESS (Livingstone and Russo, 2018) and Audio-MNIST (Becker et al., 2024) datasets, the music domain includes the MTT (Law et al., 2010) and MS-DB (Bittner et al., 2014) datasets, and the audio domain includes the ESC50 (Piczak, 2015), US8K (Salamon et al., 2014) and VIVAE (Holz et al., 2022) datasets.

As shown in Table 3, partitioning into multiple bands with separate encoders substantially enhances the semantic capacity of the encoder. For speech and audio subtasks, finer-grained frequency decomposition enables better capture of domain-specific acoustic characteristics, with the 5-band configuration achieving the best performance. For the music subtask, increasing partitions beyond a moderate level does not yield further improvements, though all band-split configurations significantly outperform the single-encoder baseline. This indicates that while band splitting is universally beneficial for semantic representation learning, the optimal granularity varies across domains, likely reflecting their distinct spectral characteristics and information distribution patterns.

6 Ablation Study

To eliminate the potential confounding effect of SimVQ itself on the experimental results, we conduct ablation studies using Residual SimVQ and replacing the VQ in band-based quantization with SimVQ. The results are presented in Table 5. As shown, neither replacing DAC's RVQ with Residual SimVQ nor substituting VQ with SimVQ in the 5-band BSCodec configuration yields significant performance improvements. In our experiments, SimVQ serves solely as a quantization tool for large codebooks. Moreover, as previously demonstrated, the codebook utilization rate of SimVQ is comparable to that of DAC. Therefore, the design of SimVQ itself is not the decisive factor contributing to the model's superior performance.

Table 5: Ablation study of VQ method.

Model	**#Band**	**Speech UTMOS↑**	**Audio VISQOL↑**	**Music VISQOL↑**
DAC	1	3.201	4.055	4.171
w/ RSimVQ	1	2.655	4.185	4.267
BSCodec	5	2.515	4.245	4.326
w/ SimVQ	5	2.723	4.185	4.267

7 Limitations

The presented BSCodec can further benefit from a more comprehensive evaluation protocol. Current evaluation concentrates on audio reconstruction 5.1 and small-scale understanding-oriented downstream tasks 5.4. Further investigation on large-scale codec-based audio generation tasks (e.g., language model-based TTS) can provide a more comprehensive profile of the strength of our BSCodec.

8 Conclusion

We present BSCodec, a band-split neural audio codec for universal audio compression that processes different frequency bands separately through parallel encoder-quantizer-decoder modules, naturally handling the spectral differences between speech, music and sound. Our experiments show that BSCodec achieves better reconstruction quality on music and sound compared to existing codecs while maintaining competitive performance on speech, and downstream task evaluations confirm that the learned representations are effective for audio understanding applications.

Acknowledgment

This work used the Bridges2 at PSC and Delta/DeltaAI NCSA systems through CIS210014 from the ACCESS program, supported by NSF #2138259, #2138286, #2138307, #2137603, and #2138296.

References

Rosana Ardila, Megan Branson, Kelly Davis, Michael Henretty, Michael Kohler, Josh Meyer, Reuben Morais, Lindsay Saunders, Francis M Tyers, and Gregor Weber. 2019. Common voice: A massively-multilingual speech corpus. *arXiv preprint arXiv:1912.06670*.

Sören Becker, Johanna Vielhaben, Marcel Ackermann, Klaus-Robert Müller, Sebastian Lapuschkin, and Wojciech Samek. 2024. Audiomnist: Exploring explainable artificial intelligence for audio analysis on a simple benchmark. *Journal of the Franklin Institute*, 361(1):418–428.

Rachel M Bittner, Justin Salamon, Mike Tierney, Matthias Mauch, Chris Cannam, and Juan Pablo Bello. 2014. Medleydb: A multitrack dataset for annotation-intensive mir research. In *Ismir*, volume 14, pages 155–160.

Dmitry Bogdanov, Minz Won, Philip Tovstogan, Alastair Porter, and Xavier Serra. 2019. The mtg-jamendo dataset for automatic music tagging. ICML.

Karlheinz Brandenburg. 1999. Mp3 and aac explained. In *Audio Engineering Society Conference: 17th International Conference: High-Quality Audio Coding*. Audio Engineering Society.

Alexandre Défossez, Jade Copet, Gabriel Synnaeve, and Yossi Adi. 2022. High fidelity neural audio compression. *arXiv preprint arXiv:2210.13438*.

Jort F Gemmeke, Daniel PW Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R Channing Moore, Manoj Plakal, and Marvin Ritter. 2017. Audio set: An ontology and human-labeled dataset for audio events. In *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 776–780. IEEE.

Ian J Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. *Advances in neural information processing systems*, 27.

Yiwei Guo, Zhihan Li, Hankun Wang, Bohan Li, Chongtian Shao, Hanglei Zhang, Chenpeng Du, Xie Chen, Shujie Liu, and Kai Yu. 2025. Recent advances in discrete speech tokens: A review. *arXiv preprint arXiv:2502.06490*.

Curtis Hawthorne, Andriy Stasyuk, Adam Roberts, Ian Simon, Cheng-Zhi Anna Huang, Sander Dieleman, Erich Elsen, Jesse Engel, and Douglas Eck. 2019. Enabling factorized piano music modeling and generation with the MAESTRO dataset. In *International Conference on Learning Representations*.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. 2022. Masked autoencoders are scalable vision learners. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15979–15988.

Andrew Hines, Jan Skoglund, Anil C Kokaram, and Naomi Harte. 2015. Visqol: an objective speech quality model. *EURASIP Journal on Audio, Speech, and Music Processing*, 2015(1):13.

Natalie Holz, Pauline Larrouy-Maestri, and David Poeppel. 2022. The variably intense vocalizations of affect and emotion (vivae) corpus prompts new perspective on nonspeech perception. *Emotion*, 22(1):213.

Po-Yao Huang, Hu Xu, Juncheng Li, Alexei Baevski, Michael Auli, Wojciech Galuba, Florian Metze, and Christoph Feichtenhofer. 2022. Masked autoencoders that listen. *Advances in Neural Information Processing Systems*, 35:28708–28720.

Shengpeng Ji, Ziyue Jiang, Wen Wang, Yifu Chen, Minghui Fang, Jialong Zuo, Qian Yang, Xize Cheng, Zehan Wang, Ruiqi Li, and 1 others. 2024. Wavtokenizer: an efficient acoustic discrete codec tokenizer for audio language modeling. *arXiv preprint arXiv:2408.16532*.

Yidi Jiang, Qian Chen, Shengpeng Ji, Yu Xi, Wen Wang, Chong Zhang, Xianghu Yue, ShiLiang Zhang, and Haizhou Li. 2025. Unicodec: Unified audio codec with single domain-adaptive codebook. *arXiv preprint arXiv:2502.20067*.

Jee-weon Jung, Wangyou Zhang, Jiatong Shi, Zakaria Aldeneh, Takuya Higuchi, Barry-John Theobald, Ahmed Hussen Abdelaziz, and Shinji Watanabe. 2024. Espnet-spk: full pipeline speaker embedding toolkit with reproducible recipes, self-supervised front-ends, and off-the-shelf models. *arXiv preprint arXiv:2401.17230*.

Rithesh Kumar, Prem Seetharaman, Alejandro Luebs, Ishaan Kumar, and Kundan Kumar. 2023. High-fidelity audio compression with improved rvqgan. *Advances in Neural Information Processing Systems*, 36:27980–27993.

Moreno La Quatra, Alkis Koudounas, Lorenzo Vaiani, Elena Baralis, Luca Cagliero, Paolo Garza, and Sabato Marco Siniscalchi. 2024. Benchmarking representations for speech, music, and acoustic events. In *2024 IEEE International Conference on Acoustics, Speech, and Signal Processing Workshops (ICASSPW)*, pages 505–509. IEEE.

Edith Law, Kris West, M Mandel, M Bay, and JS Downie. 2010. Evaluation of algorithms using

games: the case of music annotation. In *Proceedings of the 11th International Society for Music Information Retrieval Conference (ISMIR). Utrecht, the Netherlands*.

Haohe Liu, Xuenan Xu, Yi Yuan, Mengyue Wu, Wenwu Wang, and Mark D Plumbley. 2024. Semanticodec: An ultra low bitrate semantic audio codec for general sound. *IEEE Journal of Selected Topics in Signal Processing*.

Steven R Livingstone and Frank A Russo. 2018. The ryerson audio-visual database of emotional speech and song (ravdess): A dynamic, multimodal set of facial and vocal expressions in north american english. *PloS one*, 13(5):e0196391.

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Yi Luo and Jianwei Yu. 2023. Music source separation with band-split rnn. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 31:1893–1901.

Yi Luo, Jianwei Yu, Hangting Chen, Rongzhi Gu, and Chao Weng. 2024. Gull: A generative multifunctional audio codec. *arXiv preprint arXiv:2404.04947*.

Henrique S Malvar. 1990. Modulated qmf filter banks with perfect reconstruction. *Electronics letters*, 26(13):906–907.

Fabian Mentzer, David Minnen, Eirikur Agustsson, and Michael Tschannen. 2023. Finite scalar quantization: Vq-vae made simple. *arXiv preprint arXiv:2309.15505*.

Pooneh Mousavi, Gallil Maimon, Adel Moumen, Darius Petermann, Jiatong Shi, Haibin Wu, Haici Yang, Anastasia Kuznetsova, Artem Ploujnikov, Ricard Marxer, and 1 others. 2025. Discrete audio tokens: More than a survey! *arXiv preprint arXiv:2506.10274*.

Dianwen Ng, Kun Zhou, Yi-Wen Chao, Zhiwei Xiong, Bin Ma, and EngSiong Chng. 2025. Multi-band frequency reconstruction for neural psychoacoustic coding. In *Forty-second International Conference on Machine Learning*.

Karol J Piczak. 2015. Esc: Dataset for environmental sound classification. In *Proceedings of the 23rd ACM international conference on Multimedia*, pages 1015–1018.

Mathias Prokop. 2003. General principles of mdct. *European journal of radiology*, 45:S4–S10.

Zafar Rafii, Antoine Liutkus, Fabian-Robert Stöter, Stylianos Ioannis Mimilakis, and Rachel Bittner. 2017. The MUSDB18 corpus for music separation.

Antony W Rix, John G Beerends, Michael P Hollier, and Andries P Hekstra. 2001. Perceptual evaluation of speech quality (pesq)-a new method for speech quality assessment of telephone networks and codecs. In *2001 IEEE international conference on acoustics, speech, and signal processing. Proceedings (Cat. No. 01CH37221)*, volume 2, pages 749–752. IEEE.

Takaaki Saeki, Detai Xin, Wataru Nakata, Tomoki Koriyama, Shinnosuke Takamichi, and Hiroshi Saruwatari. 2022. Utmos: Utokyo-sarulab system for voicemos challenge 2022. *arXiv preprint arXiv:2204.02152*.

Justin Salamon, Christopher Jacoby, and Juan Pablo Bello. 2014. A dataset and taxonomy for urban sound research. In *Proceedings of the 22nd ACM international conference on Multimedia*, pages 1041–1044.

Tapio Saramaki and Robert Bregovic. 2002. Multirate systems and filterbanks. In *Multirate systems: design and applications*, pages 27–85. IGI Global Scientific Publishing.

Bertram Scharf. 1970. Critical bands. *Foundations of modern auditory theory*, 1:157–202.

Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. 2017. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. *arXiv preprint arXiv:1701.06538*.

Jiatong Shi, Hye-jin Shim, Jinchuan Tian, Siddhant Arora, Haibin Wu, Darius Petermann, Jia Qi Yip, You Zhang, Yuxun Tang, Wangyou Zhang, and 1 others. 2024a. Versa: A versatile evaluation toolkit for speech, audio, and music. *arXiv preprint arXiv:2412.17667*.

Jiatong Shi, Jinchuan Tian, Yihan Wu, Jee-weon Jung, Jia Qi Yip, Yoshiki Masuyama, William Chen, Yuning Wu, Yuxun Tang, Massa Baali, and 1 others. 2024b. Espnet-codec: Comprehensive training and evaluation of neural codecs for audio, music, and speech. In *2024 IEEE Spoken language technology workshop (SLT)*, pages 562–569. IEEE.

Cees H Taal, Richard C Hendriks, Richard Heusdens, and Jesper Jensen. 2010. A short-time objective intelligibility measure for time-frequency weighted noisy speech. In *2010 IEEE international conference on acoustics, speech and signal processing*, pages 4214–4217. IEEE.

Marco Tagliasacchi, Yunpeng Li, Karolis Misiunas, and Dominik Roblek. 2020. Seanet: A multi-modal speech enhancement network. *arXiv preprint arXiv:2009.02095*.

Jinchuan Tian, Jiatong Shi, William Chen, Siddhant Arora, Yoshiki Masuyama, Takashi Maekaku, Yihan Wu, Junyi Peng, Shikhar Bharadwaj, Yiwen Zhao, and 1 others. 2025. Espnet-speechlm: An open speech language model toolkit. *arXiv preprint arXiv:2502.15218*.

Chengyi Wang, Sanyuan Chen, Yu Wu, Ziqiang Zhang, Long Zhou, Shujie Liu, Zhuo Chen, Yanqing Liu, Huaming Wang, Jinyu Li, and 1 others. 2023. Neural codec language models are zero-shot text to speech synthesizers. *arXiv preprint arXiv:2301.02111*.

Shinji Watanabe, Takaaki Hori, Shigeki Karita, Tomoki Hayashi, Jiro Nishitoba, Yuya Unno, Nelson Enrique Yalta Soplin, Jahn Heymann, Matthew Wiesner, Nanxin Chen, and 1 others. 2018. Espnet: End-to-end speech processing toolkit. *arXiv preprint arXiv:1804.00015*.

Haibin Wu, Xuanjun Chen, Yi-Cheng Lin, Kaiwei Chang, Jiawei Du, Ke-Han Lu, Alexander H Liu, Ho-Lam Chung, Yuan-Kuei Wu, Dongchao Yang, and 1 others. 2024. Codec-superb@ slt 2024: A lightweight benchmark for neural audio codec models. In *2024 IEEE Spoken Language Technology Workshop (SLT)*, pages 570–577. IEEE.

Detai Xin, Xu Tan, Shinnosuke Takamichi, and Hiroshi Saruwatari. 2024. Bigcodec: Pushing the limits of low-bitrate neural speech codec. *arXiv preprint arXiv:2409.05377*.

Junichi Yamagishi, Christophe Veaux, and Kirsten MacDonald. 2019. CSTR VCTK Corpus: English multi-speaker corpus for CSTR voice cloning toolkit (version 0.92).

Dongchao Yang, Songxiang Liu, Haohan Guo, Jiankun Zhao, Yuanyuan Wang, Helin Wang, Zeqian Ju, Xubo Liu, Xueyuan Chen, Xu Tan, and 1 others. 2025. Almtokenizer: A low-bitrate and semantic-rich audio codec tokenizer for audio language modeling. *arXiv preprint arXiv:2504.10344*.

Dongchao Yang, Songxiang Liu, Rongjie Huang, Jinchuan Tian, Chao Weng, and Yuexian Zou. 2023. Hifi-codec: Group-residual vector quantization for high fidelity audio codec. *arXiv preprint arXiv:2305.02765*.

Geng Yang, Shan Yang, Kai Liu, Peng Fang, Wei Chen, and Lei Xie. 2021. Multi-band melgan: Faster waveform generation for high-quality text-to-speech. In *2021 IEEE Spoken Language Technology Workshop (SLT)*, pages 492–498. IEEE.

Zhen Ye, Peiwen Sun, Jiahe Lei, Hongzhan Lin, Xu Tan, Zheqi Dai, Qiuqiang Kong, Jianyi Chen, Jiahao Pan, Qifeng Liu, and 1 others. 2025. Codec does matter: Exploring the semantic shortcoming of codec for audio language model. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 25697–25705.

Neil Zeghidour, Alejandro Luebs, Ahmed Omran, Jan Skoglund, and Marco Tagliasacchi. 2021. Soundstream: An end-to-end neural audio codec. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30:495–507.

Heiga Zen, Viet Dang, Rob Clark, Yu Zhang, Ron J Weiss, Ye Jia, Zhifeng Chen, and Yonghui Wu. 2019. Libritts: A corpus derived from librispeech for text-to-speech. *arXiv preprint arXiv:1904.02882*.

Dengsheng Zhang. 2019. Wavelet transform. In *Fundamentals of image data mining: Analysis, Features, Classification and Retrieval*, pages 35–44. Springer.

Xin Zhang, Dong Zhang, Shimin Li, Yaqian Zhou, and Xipeng Qiu. 2023. Speechtokenizer: Unified speech tokenizer for speech large language models. *arXiv preprint arXiv:2308.16692*.

Yongxin Zhu, Bocheng Li, Yifei Xin, Zhihua Xia, and Linli Xu. 2024. Addressing representation collapse in vector quantized models with one linear layer. *arXiv preprint arXiv:2411.02038*.

A Training Convergence Speed Comparison

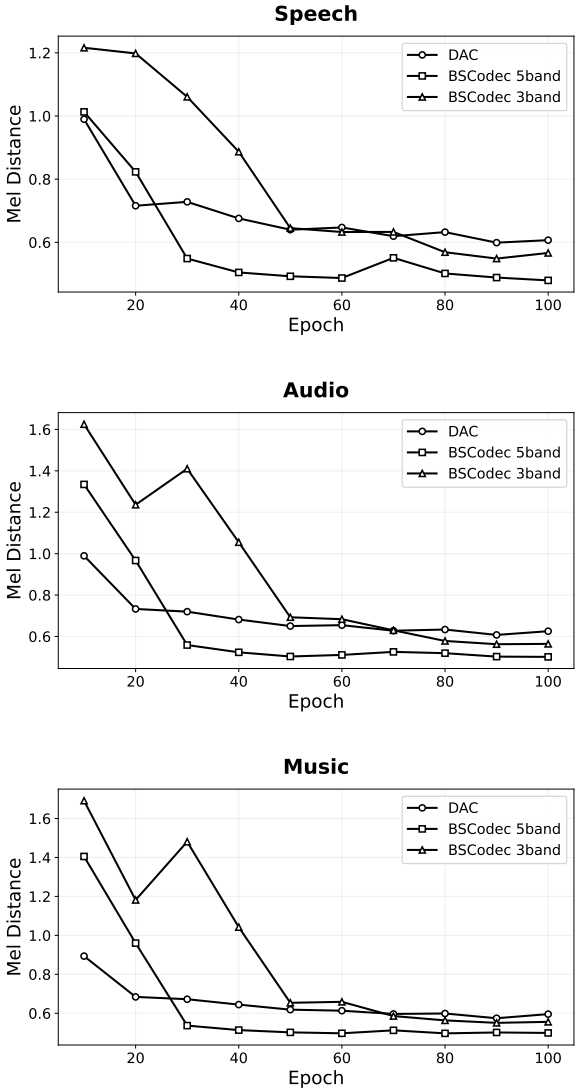

Figure 4: Comparison of the decrease in MEL distance during training for different codecs

We compare the convergence speed of different codec architectures by tracking Mel Distance throughout training on the three-domain mixed dataset. Evaluations are conducted every 10 epochs, with each epoch processing 2000 samples.

Figure 4 shows the convergence curves across the three domains. DAC exhibits slower convergence with flatter curves across all domains. Both BSCodec 5-band and 3-band configurations reach stable performance within approximately 40-60 epochs, showing steeper descent in the early training phase.

B Codebook Utilization Calculation

For a single-layer codebook with size K, the utilization rate is computed by calculating the entropy of the codebook usage distribution and comparing it to the theoretical maximum entropy. Our calculation methodology follows the approach used in DAC (Kumar et al., 2023). Let c denote the codebook index selected for a given frame, and $p(c)$ be its empirical distribution over the evaluation dataset. The single-layer entropy is:

$$H(c) = -\sum_{j=1}^{K} p(j) \log_2 p(j) \tag{7}$$

The single-layer utilization rate is then:

$$U = \frac{H(c)}{\log_2 K} \tag{8}$$

where $\log_2 K$ represents the theoretical maximum entropy for a uniform distribution over K entries.

However, this layer-wise approach fails to capture inter-layer dependencies that are critical for understanding the true capacity usage of multi-layer quantization schemes. For a codec with L layers, each having a codebook of size K, treating all layers as a single unified codebook yields a theoretical space of size K^L. For example, an 8-layer RVQ codec with $K = 1024$ per layer corresponds to a joint codebook of size 2^{80} (approximately 10^{24}). Computing the true entropy of such a massive discrete distribution would require statistics over an impractically large audio dataset and is computationally infeasible.

To measure inter-layer correlation while maintaining computational tractability, we adopt a pairwise analysis approach. Specifically, we compute the joint utilization of consecutive layer pairs, treating each pair as a combined codebook of size K^2. For layers i and $i+1$, let c_i and c_{i+1} denote the codebook indices selected for a given frame. The joint distribution is:

$$p(c_i, c_{i+1}) = \frac{\text{count}(c_i, c_{i+1})}{N} \tag{9}$$

where N is the total number of frames in the evaluation dataset. The joint entropy is:

$$H(c_i, c_{i+1}) = -\sum_{j=1}^{K} \sum_{k=1}^{K} p(j,k) \log_2 p(j,k) \tag{10}$$

The pairwise codebook utilization rate is then defined as:

$$U_{i,i+1} = \frac{H(c_i, c_{i+1})}{2 \log_2 K} \tag{11}$$

where the denominator $2 \log_2 K$ represents the theoretical maximum entropy for two independent uniform distributions over K entries. A utilization rate approaching 1.0 indicates that the two layers are nearly statistically independent and fully utilized, while values significantly below 1.0 suggest redundancy or correlation between layers.

For RVQ-based codecs, we compute $U_{i,i+1}$ for all consecutive pairs $(i, i+1)$ where $i \in \{1, 2, \ldots, L-1\}$ to assess the degree of independence across the residual hierarchy.

C Speech Reconstruction Metrics

We provide detailed descriptions of the speech quality metrics used in our evaluation.

Wide-Band Perceptual Evaluation of Speech Quality (WB-PESQ) is an ITU-T P.862.2 standard metric that predicts subjective listening quality by comparing processed speech to its clean reference. It outputs scores ranging from -0.5 to 4.5, with higher values indicating better perceptual quality.

Short-Time Objective Intelligibility (STOI) evaluates speech intelligibility by measuring the correlation between short-time temporal envelopes of processed and reference signals. Scores range from 0 to 1, with higher values indicating better intelligibility.

UTMOS (University of Tokyo Mean Opinion Score) is a deep learning-based metric that predicts the subjective Mean Opinion Score of synthesized speech without human listeners. It provides scores from 1 to 5, where higher scores indicate better perceived quality.

SPK_SIM (Speaker Similarity) measures the preservation of speaker characteristics by computing the cosine similarity between speaker embeddings extracted from synthesized and ground-truth utterances using ESPnet-SPK (Jung et al., 2024). Values range from 0 to 1, with higher scores indicating better speaker identity preservation.

VISQOL (Virtual Speech Quality Objective Listener) is a full-reference metric that predicts perceived audio quality by measuring spectro-temporal similarity between processed and reference signals. Scores range from 1 to 5, with higher values indicating better quality.

D Energy Distribution Extraction

We extract energy distributions from audio files using the following pipeline:

Preprocessing: All audio samples are normalized to -23.0 LUFS using pyloudnorm to ensure fair comparison across recordings.

Spectral analysis: We compute the Short-Time Fourier Transform (STFT) with FFT size $N_{\text{FFT}} = 2048$ and hop length $H = 512$. The energy spectrum is obtained as $E_i(f,t) = |X_i(f,t)|^2$, where $X_i(f,t)$ are the STFT coefficients for file i.

Weighted averaging: For each category (sound, music, speech), we compute the weighted average across all files:

$$E_{\text{category}}(f) = \frac{\sum_{i=1}^{N} \sum_{t=1}^{T_i} E_i(f,t)}{\sum_{i=1}^{N} T_i} \tag{12}$$

where N is the number of files and T_i is the number of frames in file i. The results are plotted on a semi-log scale for 0–8 kHz.

E Data Distribution

Table 6: Datasets of Stage 1.

Dataset	Duration (hours)	Description
MTG-Jamendo	3,764.42	Open music dataset
MAESTRO	194.27	Classical piano performances with aligned MIDI
MUSDB18	6.04	150 full-length music tracks
Total	3,964.73	

Table 7: Datasets of Stage 2.

Dataset	Duration (hours)	Description
AudioSet	742.51	Large-scale manually annotated audio events
CommonVoice	14.93	Multilingual transcribed speech corpus
MTG-Jamendo	753.82	Open music dataset
LibriTTS	528.94	Multi-speaker English audiobook corpus
MUSDB18	6.04	150 full-length music tracks
VCTK	80.80	110 English speakers with various accents
Total	2,127.04	