SACodec: Asymmetric Quantization with Semantic Anchoring for Low-Bitrate High-Fidelity Neural Speech Codecs

Zhongren Dong[1], Bin Wang[2], Jing Han[1]*, Haotian Guo[1], Xiaojun Mo[1], Yimin Cao[1], Zixing Zhang[1,3]

[1]College of Computer Science and Electronic Engineering, Hunan University, Changsha, China
[2]Beijing Xiaomi Mobile Software Co., Ltd, Beijing, China
[3]Shenzhen Research Institute, Hunan University, Shenzhen, China
jhan@hnu.edu.cn

Abstract

Neural Speech Codecs face a fundamental trade-off at low bitrates: preserving acoustic fidelity often compromises semantic richness. To address this, we introduce SACodec, a novel codec built upon an asymmetric dual-quantizer that employs our proposed **S**emantic **A**nchoring mechanism. This design strategically decouples the quantization of **S**emantic and **A**coustic details. The semantic anchoring is achieved via a lightweight projector that aligns acoustic features with a frozen, large-scale mHuBERT codebook, injecting linguistic priors while guaranteeing full codebook utilization. Sequentially, for acoustic details, a residual activation module with SimVQ enables a single-layer quantizer (acoustic path) to faithfully recover fine-grained information. At just 1.5 kbps, SACodec establishes a new state of the art by excelling in both fidelity and semantics: subjective listening tests confirm that its reconstruction quality is perceptually highly comparable to ground-truth audio, while its tokens demonstrate substantially improved semantic richness in downstream tasks.

Code — https://github.com/SmileHnu/SACodec

Introduction

The deepening integration of Large Language Models (LLMs) into the speech domain has made Neural Speech Codecs (NSCs)—the crucial bridge between continuous waveforms and discrete tokens—increasingly pivotal (Défossez et al. 2024; Xu et al. 2025; Ding et al. 2025). By converting high-dimensional signals into low-dimensional symbol sequences, discrete tokens form the bedrock of modern Speech Language Models (SLMs), enabling powerful LLM architectures to be applied directly to tasks like Text-to-Speech (TTS) synthesis (Chen et al. 2025, 2024; Du et al. 2025) and spoken dialogue (Borsos et al. 2023; Zhang et al. 2023; Ma et al. 2025). Consequently, developing low-bitrate codecs has emerged as a central research frontier. Traditionally, low bitrates played a role in reducing communication/storage costs; in the LLM era, their value has expanded to enhancing computational efficiency in large-scale models. Efficient codecs mitigate the quadratic complexity of attention mechanisms by generating shorter token sequences, thereby reducing inference latency and cost, which is crucial for real-time audio-language model services (Li et al. 2024; Xin et al. 2024).

Yet, the pursuit of ever-lower bitrates exposes a fundamental trade-off in high-fidelity codecs like Encodec (Défossez et al. 2023) and DAC (Kumar et al. 2023), which rely on multi-layer Residual Vector Quantization (RVQ) (Zeghidour et al. 2021). While effective across a range of bitrates, RVQ's performance degrades sharply when the bitrate budget is constrained to 1.5 kbps. The accumulation of quantization errors across layers yields audible artifacts (Guo et al. 2025), and more critically, the resulting multi-stream tokens introduce significant downstream modeling complexity. SLMs now employ intricate parallel or non-autoregressive decoders to handle these hierarchical token streams (Chen et al. 2025; Borsos et al. 2023), a stark mismatch with the community's drive for the architectural simplicity of a single, unified sequence.

To mitigate this modeling complexity, recent works led by WavTokenizer (Ji et al. 2025) have pioneered the single-codebook paradigm. By compressing all information into a single discrete sequence, these codecs greatly simplify downstream integration, enabling a more direct application of standard language modeling. However, this acoustics-oriented optimization, driven solely by signal-distortion objectives, yields representations that lack explicit semantic structure. This becomes a liability for tasks requiring deep content understanding, shifting the challenge from "how to compress efficiently" to the more profound question of "how to encode meaningful content" (Zhang et al. 2024a).

To infuse codecs with semantics, the community has explored two dominant strategies: external knowledge distillation (Zhang et al. 2024a; Défossez et al. 2024) and endogenous self-supervised learning (Jiang et al. 2025). Both routes, however, lead to an uneasy compromise. To preserve reconstruction quality, these methods either retain a complex multi-layer RVQ backend, conflicting with the low-bitrate objective, or introduce significant training complexity and computational overhead. Fundamentally, to endow tokens with semantic capacity, the selected codebooks from the quantizer must carry semantic information, rather than serving merely for audio reconstruction. However, traditional VQ is inherently inefficient due to its local update

*Corresponding author.

rule—only the nearest codebook is updated—leading to codebook collapse and severely limiting both the expressivity and scalability of the learned codebook (Esser, Rombach, and Ommer 2021; Zhu et al. 2024a,b). Breaking this stalemate thus demands a new quantization paradigm. Recent advancements like SimVQ (Zhu et al. 2024b), which reparameterize the codebook to enable global updates, have shown a promising path toward achieving high codebook utilization. A key challenge, which we address, is how to seamlessly couple such efficient quantization with a direct mechanism for semantic injection.

To overcome the twin bottlenecks of VQ inefficiency and cumbersome semantic injection, we introduce a novel **S**emantic-**A**nchored speech codec, namely **SACodec**. Our approach is founded on a *asymmetric dual-quantizer* architecture that assigns specialized, highly efficient quantization mechanisms to distinct semantic and acoustic information streams. By strategically decoupling the modeling of this information at the quantization level, we address the core trade-offs that constrain existing codecs. Our contributions are threefold:

- We propose a novel semantic anchoring mechanism that leverages a fixed, large-scale mHuBERT codebook. A lightweight learned projector adaptively aligns this external knowledge to the acoustic latent space, efficiently injecting strong semantic priors while effectively preventing codebook collapse in the semantic layer.
- We introduce a synergistic residual activation module that equips a single-layer VQ with the SimVQ technique. This design guarantees full codebook activation for the residual quantizer, enabling it to compensate for fine-grained acoustic details with minimal architectural complexity and bitrate overhead.
- We demonstrate through extensive experiments that SACodec establishes a new record for low-bitrate speech codecs. At 1.5 kbps, it achieves superior reconstruction quality over all baselines and exhibits stronger semantic representation in downstream tasks, offering a better-balanced solution for modern SLMs.

Related Works

Recent advances in NSCs have evolved along three main dimensions: acoustic codec paradigms, semantic enhancement strategies, and training paradigms. We review each to contextualize SACodec's contributions.

Acoustic Codec Paradigms. Dominant high-fidelity NSCs (SoundStream (Zeghidour et al. 2021), Encodec (Défossez et al. 2023), DAC (Kumar et al. 2023), HiFi-Codec (Yang et al. 2023)) rely on RVQ. By cascading multiple codebooks, RVQ-based models excel at reconstruction but produce multi-stream tokens, which complicates downstream autoregressive modeling (Guo et al. 2025). Recent single-codebook codecs (WavTokenizer (Ji et al. 2025), BigCodec (Xin et al. 2024), Single-Codec (Li et al. 2024)) output single token sequences, simplifying integration with language models and enabling efficient compression. However, a fundamental challenge shared by both multi-layer and single-layer approaches is codebook collapse, where only a fraction of the learnable codebook is utilized, capping their ultimate representational power (Esser, Rombach, and Ommer 2021; Zhu et al. 2024a,b).

Semantic Enhancement Strategies. To address the semantic sparsity of acoustic-only codecs, two main strategies have emerged. The first relies on external knowledge distillation. Models like SpeechTokenizer (Zhang et al. 2024a) and Mimi (Défossez et al. 2024) use pre-trained SSL models (e.g., HuBERT (Hsu et al. 2021), WavLM (Chen et al. 2022)) as "teachers" to guide early quantization. This reflects a broader trend of injecting semantic priors, with some works even employing fixed semantic codebooks from text models like LLaMA (Yang et al. 2024). While effective, this strategy introduces a significant dependency on external, often large-scale, pre-trained models, and still requires a multi-layer RVQ backend to compensate for the potential degradation in reconstruction quality, which conflicts with low-bitrate goals.

The second route pursues endogenous semantic learning. This includes methods ranging from disentangling speech attributes via multi-task supervision in FACodec (Ju et al. 2024) to decoupling speaker timbre in LSCodec (Guo et al. 2024). A parallel line of work also explores diffusion-based models like SemantiCodec (Liu et al. 2024), which generate acoustics from semantic tokens. These approaches, while more self-contained, often lead to a different set of trade-offs: the disentanglement process can be fragile and hard to optimize, while diffusion-based decoders introduce substantial computational overhead during inference, limiting their applicability in real-time scenarios.

Training Paradigms and the Unifying Dilemma. From a training perspective, the dual objectives of acoustic fidelity and semantic richness have led to complex paradigms. Many advanced codecs adopt staged or intricate multi-task frameworks to balance these competing goals. For instance, some approaches rely on multi-stage pipelines (Liu et al. 2024), while others require separate modules for semantic and acoustic modeling (Ju et al. 2024). Although effective, such non-monolithic designs often hinder reproducibility and complicate optimization, as highlighted in recent benchmarks (Mousavi et al. 2025). Conversely, while simple end-to-end paradigms are attractive, they have traditionally struggled to integrate semantic information effectively. Even in tasks like TTS, bridging the gap between textual and acoustic tokens frequently requires complex cascaded models (Kharitonov et al. 2023).

These challenges stem from a common root: the inefficiency of conventional VQ. This forces a compromise among architectural simplicity, direct semantic injection, and low-bitrate performance. SACodec confronts this unifying bottleneck directly. By integrating two distinct, highly efficient quantization mechanisms—one for anchoring semantics and one for activating residuals—within a single end-to-end framework, our work offers a novel approach to this long-standing trilemma.

Methodology

This section details SACodec, a novel neural speech codec for low-bitrate, high-fidelity, and semantically rich tokeniza-

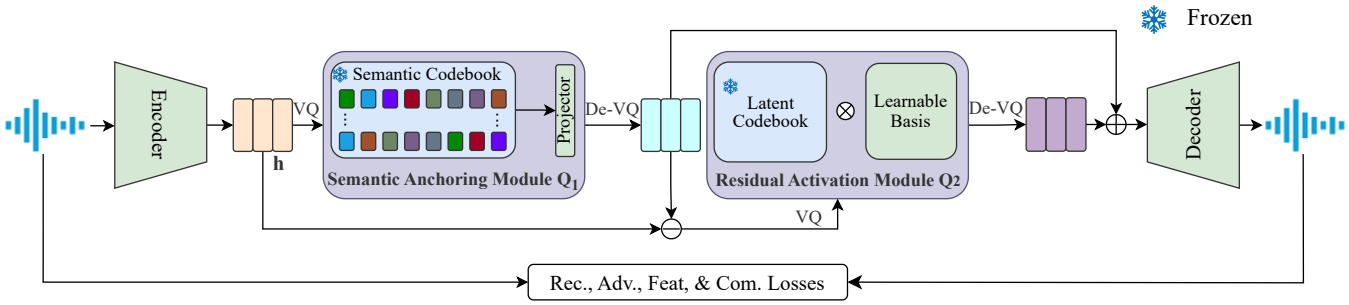

Figure 1: The architecture of SACodec, centered on our **Asymmetric Dual Quantizer**. An input waveform is mapped by a convolutional-recurrent **Encoder** to a latent representation $\mathbf{h}$. This representation is processed sequentially by two specialized quantization modules: **(1) Semantic Anchoring (Q1):** A learnable Projector aligns $\mathbf{h}$ with a large-scale, pre-computed, and frozen Semantic Codebook derived from mHuBERT. This process anchors the core linguistic content. **(2) Residual Activation (Q2):** The acoustic residual ($\mathbf{h}$ minus the semantic embedding) is quantized by a single-layer VQ. This quantizer employs the SimVQ technique, where a Learnable Basis transforms a randomly initialized and frozen Latent Codebook to dynamically form the residual space, ensuring full codebook activation. During training, only the Projector and the Learnable Basis are updated. The outputs of both quantizers are summed and passed to ConvNeXt-Attention **Decoder**, which reconstructs the speech signal via an iSTFT.

tion. We first present the overall architecture, then dissect its core innovation—the Asymmetric Dual Quantizer—and finally formulate the unified training objective.

Finally, the quantized embeddings from both modules, $\mathbf{e}_1$ and $\mathbf{e}_2$, are fused by element-wise addition to form the final representation $\mathbf{e}_{\text{final}} = \mathbf{e}_1 + \mathbf{e}_2$. This is passed to the decoder to reconstruct the high-fidelity waveform $\hat{\mathbf{x}}$. The entire generator is trained adversarially against an ensemble of multi-scale and multi-period discriminators.

Overall Framework

SACodec is built upon a GAN-based (Goodfellow et al. 2014), end-to-end framework, with its full pipeline illustrated in Fig. 1. The model comprises three core components: an encoder, our proposed asymmetric dual quantizer, and a decoder, which operate sequentially.

First, the **encoder** processes an input speech waveform $\mathbf{x} \in \mathbb{R}^L$. Following the Encodec backbone (Défossez et al. 2023), it employs a convolutional stack with ELU activations and a two-layer LSTM (Hochreiter and Schmidhuber 1997). The encoder processes a 24 kHz waveform through a series of strided convolutions, achieving a 320× downsampling to a 75 Hz frame rate. A final linear layer then projects the features to the target dimension D, producing the continuous latent representation $\mathbf{h} \in \mathbb{R}^{T \times D}$.

Next, this latent representation $\mathbf{h}$ is processed by our **asymmetric dual quantizer**. It operates in two stages: the Semantic Anchoring Module $\mathbf{Q}_1$ first extracts the core semantic content by quantizing $\mathbf{h}$ against a projected mHuBERT codebook, yielding an embedding $\mathbf{e}_1$. The resulting acoustic residual is then quantized by the Residual Activation Module $\mathbf{Q}_2$, which uses the SimVQ technique to capture fine-grained acoustic details in a second embedding, $\mathbf{e}_2$.

Finally, the quantized embeddings from both modules are fused by element-wise addition ($\mathbf{e}_{\text{final}} = \mathbf{e}_1 + \mathbf{e}_2$) and passed to the **decoder**. Our decoder is engineered for high-fidelity synthesis from this fused representation. Inspired by modern vocoders (Ji et al. 2025; Siuzdak 2024), it decouples feature processing from signal synthesis. A powerful ConvNeXt-Attention backbone first models both local and global dependencies in the feature sequence. The resulting features are then projected to a complex spectrogram and deterministically converted to the output waveform $\hat{\mathbf{x}}$ via an inverse Short-Time Fourier Transform (iSTFT).

The entire generator (encoder, quantizer, and decoder) is trained adversarially against an ensemble of multi-scale and multi-period discriminators.

Asymmetric Dual Quantizer

Semantic Anchoring Module To overcome the codebook collapse that plagues traditional learnable VQ (Zhu et al. 2024a,b) and to directly inject strong semantic priors, our semantic quantizer ($\mathbf{Q}_1$) is built upon a fixed external knowledge base. Note, we adopt the publicly available semantic codebook $\mathbf{C}_{\text{sem}} \in \mathbb{R}^{K_1 \times D_s}$, which consists of K_1=1000 centroids clustered from mHuBERT features (Lee et al. 2022), serving as a stable "semantic anchor" for our model.

To bridge the distributional gap between the encoder's acoustic representation $\mathbf{h}$ and the fixed semantic space, we adopt a codebook-space projection strategy (Zhu et al. 2024a). We learn a lightweight linear projector $\mathbf{P}_{\text{sem}}$ that transforms the *entire* frozen codebook $\mathbf{C}_{\text{sem}}$ into a dynamically adapted, effective codebook, which we denote as $\mathcal{C}_1$:

$$\mathcal{C}_1 = \mathbf{P}_{\text{sem}}(\mathbf{C}_{\text{sem}}), \tag{1}$$

where $\mathbf{P}_{\text{sem}}$ maps the source codebook into the encoder's latent space of dimension D. For each frame $\mathbf{h}_t$, the quantization index i_t and embedding $\mathbf{e}_{1,t}$ are found via nearest-neighbor lookup in this adapted codebook:

$$i_t = \arg\min_k \|\mathbf{h}_t - \mathbf{c}_{1,k}\|_2^2, \quad \text{where } \mathbf{c}_{1,k} \in \mathcal{C}_1. \tag{2}$$

This global transformation is designed to encourage full codebook utilization and enhance reconstruction quality, as further examined in our ablation studies (Fig. 3).

Residual Activation Module The semantic embedding $\mathbf{e}_{1,t}$ captures content but discards perceptually crucial acoustic details. We define this information, which includes vital paralinguistic attributes such as speaker timbre, prosodic rhythm, and speaking style, as the acoustic residual $\mathbf{r}_t$:

$$\mathbf{r}_t = \mathbf{h}_t - \mathbf{e}_{1,t}. \tag{3}$$

Reliably representing this residual is critical. Recent benchmarks (Zhang et al. 2024b) and representation learning methods (Zhang et al. 2025b) have underscored that these rich paralinguistic cues are not merely acoustic artifacts but are central to a vast range of downstream tasks, such as speaker verification and emotion recognition.

To efficiently quantize this residual, our second quantizer module, $\mathbf{Q}_2$, uses a single-layer vector quantizer enhanced by SimVQ (Zhu et al. 2024b). Instead of learning a residual codebook directly, SimVQ reparameterizes it as the product of a frozen, randomly initialized coefficient matrix $\mathbf{C}_{\text{coeff}} \in \mathbb{R}^{K_2 \times d}$ and a learnable linear "latent basis" $\mathbf{W}_{\text{basis}} \in \mathbb{R}^{d \times D}$:

$$\mathcal{C}_2 = \mathbf{C}_{\text{coeff}} \times \mathbf{W}_{\text{basis}}. \tag{4}$$

During training, only $\mathbf{W}_{\text{basis}}$ is updated. Gradients flow back to this shared basis, globally updating the entire effective residual codebook $\mathcal{C}_2$. This guarantees full codebook activation for the K_2=1024 entries. The quantized residual embedding $\mathbf{e}_{2,t}$ is then found via nearest-neighbor lookup in $\mathcal{C}_2$. This design provides a highly efficient acoustic detail encoder at minimal architectural cost.

Training Objective

SACodec is trained end-to-end within a GAN framework. The generator G (the codec itself) is optimized via a composite loss function designed to balance reconstruction fidelity, perceptual quality, and quantization stability, while a set of discriminators $\{D_k\}$ is trained to distinguish real from generated audio.

Our overall generator loss $\mathcal{L}_G$ is a weighted sum of several standard components widely used in high-fidelity speech synthesis (Kong, Kim, and Bae 2020; Kumar et al. 2023):

$$\mathcal{L}_G = \lambda_{\text{rec}}\mathcal{L}_{\text{rec}} + \lambda_{\text{adv}}\mathcal{L}_{\text{adv}} + \lambda_{\text{feat}}\mathcal{L}_{\text{feat}} + \lambda_{c1}\mathcal{L}_{\text{com},1} + \lambda_{c2}\mathcal{L}_{\text{com},2}. \tag{5}$$

These components include: (1) a multi-scale mel-spectrogram reconstruction loss ($\mathcal{L}_{\text{rec}}$) for spectral accuracy; (2) an adversarial loss ($\mathcal{L}_{\text{adv}}$) based on a powerful ensemble of multi-period (MPD) and multi-band multi-scale STFT (MS-STFT) discriminators to enhance perceptual realism; (3) a feature-matching loss ($\mathcal{L}_{\text{feat}}$) to stabilize GAN training; and (4) two commitment losses ($\mathcal{L}_{\text{com},1}$ and $\mathcal{L}_{\text{com},2}$) to regularize the encoder's outputs for the semantic and residual modules, respectively. The weights for the reconstruction, adversarial, and feature-matching losses are set to $\lambda_{\text{rec}} = 45.0$, $\lambda_{\text{adv}} = 1.0$, and $\lambda_{\text{feat}} = 1.0$ respectively, consistent with standard practices in WavTokenizer (Ji et al. 2025). The commitment losses are weighted asymmetrically ($\lambda_{c1} = 25.0, \lambda_{c2} = 5.0$). A stronger weight on the semantic branch (λ_{c1}) is necessary to enforce the alignment of encoder features with the fixed, external mHuBERT space, while the dynamically learned residual branch requires a weaker regularization.

Experimental Setup

Datasets and Baselines. Our model is trained on the 585-hour LibriTTS corpus (Zen et al. 2019), using randomly cropped 1-second segments of 24 kHz audio. We evaluate performance across three conditions: in-domain clean (*LibriTTS test-clean*) and noisy (*test-other*) splits to assess robustness, and on the out-of-domain dataset (*LJSpeech*) (Pratap et al. 2020) to measure generalization. Our model is benchmarked against a comprehensive suite of SOTA codecs, including RVQ-based models (**Encodec**, **DAC**), semantic-distilled (**SpeechTokenizer**), disentanglement-focused (**FACodec**), and single-codebook (**WavTokenizer**) approaches. For fairness, all baselines are set to operate at 1.5 kbps whenever possible.

Implementation and Training. Our model, SACodec, is built in PyTorch. It employs an Encodec-style encoder with an LSTM and a modern ConvNeXt-Attention decoder inspired by WavTokenizer. The core asymmetric dual quantizer is configured with a 1000-entry fixed semantic codebook (K_1) and a 1024-entry SimVQ-activated residual codebook (K_2). The model is trained end-to-end using the AdamW optimizer.

Evaluation Metrics. Our evaluation is twofold. **Reconstruction Quality** is measured using objective metrics—UTMOS, PESQ, STOI, and V/UV F1—and complemented by subjective MUSHRA listening tests (Series 2014). **Semantic Capability** is assessed using the comprehensive ARCH benchmark (Quatra et al. 2024), which comprises multiple downstream classification tasks across speech, music, and audio domains. We evaluate this at two levels: (1) *Compressed Domain*, which tests the intrinsic semantic richness of the raw tokens, and (2) *Reconstruction Domain*, a novel evaluation we introduce to measure end-to-end semantic fidelity by analyzing the reconstructed waveform. This second dimension is crucial as it reveals whether semantic information survives the full generation pipeline.

Results and Analysis

Main Results

Acoustic Reconstruction Quality. As shown in Table 1, SACodec establishes a new performance record in reconstruction quality for codecs operating at 1.5 kbps, showing robust performance across diverse evaluation conditions.

On in-domain clean speech (*LibriTTS test-clean*), our model's superiority at 1.5 kbps is unequivocal. It achieves a UTMOS of 4.0373, over 2.5x higher than Encodec (1.5551) and 2x higher than DAC (1.9152) at 1.5 kbps. Furthermore, it surpasses the aggressively compressed 0.9 kbps WavTokenizer, confirming that our architecture effectively translates a modest increase in bitrate into substantial fidelity gains.

The model's robustness is particularly evident on the challenging, noisy *LibriTTS test-other* set. Here, SACodec's UTMOS of 3.4786 is not only the highest among all low-bitrate competitors but is also nearly identical to the ground-truth audio's score of 3.483, indicating exceptional performance in reverberant conditions. This level of fidelity is remarkable, as it also exceeds that of higher-bitrate models, includ-

Dataset	Model	Params [M]	Token Rate	Codebook Size	Q	Bitrate [kbps]	UTMOS↑	PESQ↑	STOI↑	F1↑
LibriTTS Test-clean	GT	-	-	-	-	-	4.0562	-	-	-
	DAC	74.71	75	1024	8	6	3.6905	3.5215	.9546	.9710
	Encodec	14.85	75	1024	8	6	3.0399	2.7202	.9391	.9527
	SpeechTokenizer	103.68	50	1024	8	4	3.8794	2.6121	.9165	.9495
	FACodec	374.49	80	1024	6	4.8	3.4454	2.2532	.9127	.9402
	DAC	74.71	75	1024	2	1.5	1.9152	1.5300	.8453	.8957
	Encodec	14.85	75	1024	2	1.5	1.5551	1.5398	.8462	.8496
	SpeechTokenizer	103.68	50	1024	3	1.5	2.4121	1.2668	.7853	.8353
	WavTokenizer	80.55	75	4096	1	0.9	3.9687	2.4687	.9194	**.9394**
	SACodec(Ours)	75.17	75	1000 / 1024	2	1.5	**4.0373**	**2.6937**	**.9317**	.9381
LibriTTS Test-other	GT	-	-	-	-	-	3.4831	-	-	-
	DAC	74.71	75	1024	8	6	3.1338	3.3429	.9402	.9598
	Encodec	14.85	75	1024	8	6	2.6568	2.6818	.9241	.9338
	SpeechTokenizer	103.68	50	1024	8	4	3.2851	2.3269	.8811	.9205
	FACodec	374.49	80	1024	6	4.8	2.9302	2.0321	.8832	.9080
	DAC	74.71	75	1024	2	1.5	1.7443	1.5039	.8218	.8636
	Encodec	14.85	75	1024	2	1.5	1.5132	1.5753	.8291	.8228
	SpeechTokenizer	103.68	50	1024	3	1.5	2.0104	1.2241	.7780	.7445
	WavTokenizer	80.55	75	4096	1	0.9	3.4315	2.2705	**.9173**	.8907
	SACodec(Ours)	75.17	75	1000 / 1024	2	1.5	**3.4786**	**2.4016**	.9040	**.9273**
LJSpeech	GT	-	-	-	-	-	4.3794	-	-	-
	DAC	74.71	75	1024	8	6	4.0415	3.4294	.9567	.9670
	Encodec	14.85	75	1024	8	6	3.2281	2.6636	.9442	.9555
	SpeechTokenizer	103.68	50	1024	8	4	4.2371	2.6411	.9346	.9453
	FACodec	374.49	80	1024	6	4.8	3.9760	2.3234	.9220	.9338
	DAC	74.71	75	1024	2	1.5	1.8169	1.4307	.8487	.8904
	Encodec	14.85	75	1024	2	1.5	2.3900	2.0195	.9058	**.9326**
	SpeechTokenizer	103.68	50	1024	3	1.5	3.5281	1.6965	.8790	.9154
	WavTokenizer	80.55	75	4096	1	0.9	3.8755	1.9532	.9007	.9106
	SACodec(Ours)	75.17	75	1000 / 1024	2	1.5	**3.9912**	**2.4249**	**.9224**	.9302

Table 1: Objective reconstruction results across multiple datasets. We evaluate performance on **in-domain** clean (*LibriTTS-clean*) and noisy (*LibriTTS-other*) conditions, as well as on **out-of-domain** data (*LJSpeech*) to assess generalization. Best performance in the low-bitrate (1.5 kbps) regime is highlighted in **bold** and the second best is underlined. Q: #quantizers.

ing the 6 kbps DAC (3.1338), the 4 kbps SpeechTokenizer (3.2851), and even the much larger FaCodec (2.930).

To assess generalization, we evaluated the models on the out-of-domain *LJSpeech* dataset. SACodec performs strongly with a UTMOS of 3.9912, surpassing the single-codebook WavTokenizer and remaining competitive with higher-bitrate models like the 6 kbps DAC (4.0415). This demonstrates that our architecture generalizes well beyond the training corpus to new acoustic domains.

Subjective Evaluation To directly assess perceptual quality, we conducted a MUSHRA subjective listening test, with the results visualized as box plots in Fig. 2.The results affirm SACodec's near-ground-truth quality. At just 1.5 kbps, SACodec achieves a high median score of 96.8, placing it in the same perceptual tier as the ground-truth audio's median of 97.5 and indicating consistently high quality. This stands in stark contrast to the low-bitrate DAC, which registered

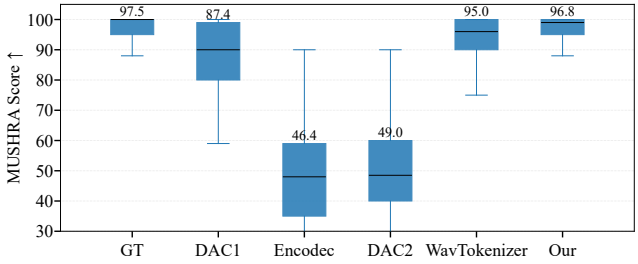

Figure 2: MUSHRA subjective evaluation on LibriTTS test-clean. Box plots show the distribution of listener scores (0-100, higher is better). The number above each box indicates the median score. DAC1: 6 kbps, DAC2: 1.5 kbps.

a low median of just 49.0 alongside a wide score distribution, signifying poor and inconsistent quality. Furthermore,

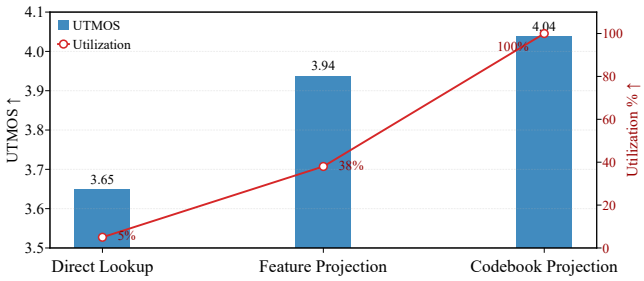

Figure 3: Comparison of semantic anchoring strategies. Our proposed Codebook Projection achieves near-perfect codebook utilization and significantly outperforms alternative strategies in reconstruction quality (UTMOS).

SACodec maintains a clear performance edge over both the highly-regarded WavTokenizer and the 6 kbps DAC baseline, the latter of which exhibits greater score variance while only reaching a median of 87.4. These results offer strong subjective evidence that our architecture produces speech with high naturalness and few audible artifacts.

Semantic Representation Richness The results on the ARCH benchmark, summarized in Table 2, highlight the strong semantic capabilities of SACodec. Notably, SACodec and SpeechTokenizer were trained solely on speech data, while for fair comparison, we evaluate WavTokenizer using its speech-only trained model, despite its availability with multi-domain (speech, music, audio) training. In contrast, DAC and Encodec were trained on large-scale, multi-domain corpora.

In *compressed domain*, this context makes SACodec's performance even more impressive. Its mean accuracy of 0.4809 considerably outperforms the speech-only WavTokenizer (0.3816) and, remarkably, remains highly competitive with the multi-domain trained DAC (0.4332), especially in the speech-related tasks. This validates that our semantic anchoring strategy effectively injects rich, generalizable semantic information.

In *reconstruction domain*, SACodec demonstrates robust end-to-end semantic fidelity. Its mean score of 0.6311 not only surpasses the speech-only WavTokenizer but also achieves parity with the 6 kbps DAC, which was trained on a much larger dataset. Crucially, SACodec's key advantage is its high consistency between the compressed and reconstruction domains. In contrast, SpeechTokenizer—despite achieving the highest mean score in the compressed domain (0.5099)—fails to maintain this lead after low-bitrate reconstruction. This discrepancy suggests a potential pitfall in its decoder: an inability to fully preserve the intrinsic semantics of its tokens. SACodec successfully avoids this pitfall. Its strong semantic fidelity even extends to unseen music and audio domains, underscoring the robustness of its asymmetric architecture. While the multi-domain trained DAC maintains an edge on the music-specific MS-DB task, SACodec's superior performance on core speech-related tasks ultimately validates the effectiveness of our semantic anchoring strategy.

Ablation Study

We conducted a series of targeted ablation studies to validate our asymmetric dual-quantizer architecture, with full results in Fig. 3 and Table 3.

Our semantic anchoring strategy is critical for both utilization and quality. As shown in Fig. 3, a naive direct lookup between the encoder and the fixed semantic codebook results in a near-zero 5% codebook utilization and a low UTMOS of 3.65. Our proposed codebook-space projection resolves this, achieving nearly 100% utilization and boosting the UTMOS to 4.04, confirming its superiority over intermediate strategies like feature projection.

Each quantizer plays a distinct and indispensable role. As shown in Table 3, the *semantic anchor* ($\mathbf{Q}_1$) is unequivocally the source of semantic representation; removing it ('w/o Q1') or replacing its pre-trained knowledge with a standard learnable VQ ('w/ Random') causes semantic accuracy to collapse by up to 30%. In parallel, the *residual activator* ($\mathbf{Q}_2$) is essential for acoustic fidelity. Removing it ('w/o Q2') severely degrades reconstruction quality, causing a 12.8% plunge in the PESQ score. Interestingly, this "semantic-only" configuration yields the highest semantic score (0.5065), highlighting an inherent tension between acoustic detail and semantic purity that our full model successfully reconciles through its synergistic design. Furthermore, the residual codebook size ablation validates our design choice for bitrate efficiency. As Table 3 shows, doubling the residual codebook size to 2048 ('w/ Larger Residual') offers only negligible gains in reconstruction quality (e.g., UTMOS increases from 4.0373 to 4.0402), confirming that our chosen size of K_2=1024 achieves a superior performance-to-bitrate trade-off.

Discussion

Training and Data Efficiency. SACodec exhibits significant training advantages. On the same 585h LibriTTS dataset, it achieves a >6x training speedup per epoch over WavTokenizer due to its architectural aptitude for shorter audio chunks. Furthermore, it reaches SOTA performance on this public dataset, while baselines like Encodec, DAC and FACodec require massive proprietary corpora. All these highlight our model's superior data and computation efficiency.

Limitations and Future Work. Our study is confined to English; future work should evaluate cross-lingual robustness. While token-level evaluations indicate effective performance, direct integration into downstream SLMs and TTS systems is required for ultimate validation. Finally, scaling the semantic codebook and exploring model compression for on-device deployment are promising future directions (Ding et al. 2021; Zhang et al. 2025a).

Conclusion

This paper introduced Semantic-Anchored speech codec (SACodec), a neural speech codec that addresses the core trade-off between acoustic quality and semantic richness at low bitrates. At a mere 1.5 kbps, SACodec delivers reconstruction quality comparable to ground-truth audio while

	Model	Token Rate	Q	Bitrate [kbps]	Speech↑		Music↑		Audio↑		$Avg.$↑
					RAVDESS	AM	MTT	MS-DB	ESC50	VIVAE	
Compressed	Encodec	75	8	6	.3507	.4913	.3097	.4226	.2015	.2675	.3406
	DAC	75	8	6	.3889	.7295	.3331	.5948	.2440	.3077	.4330
	SpeechTokenizer	50	8	4	.4896	.9725	.3591	.5566	.3790	.3027	.5099
	Encodec	75	2	1.5	.2778	.6135	.2918	.5001	.2695	.3079	.3768
	DAC	75	2	1.5	.4236	.6840	.3023	**.5859**	.2865	.3166	.4332
	FACodec	80	1	0.8	.4231	.8297	.2766	.5364	**.3665**	.3311	.4606
	WavTokenizer	75	1	0.9	.3438	.6292	.2689	.5322	.2340	.2815	.3816
	SACodec(Ours)	75	1	0.75	**.4265**	**.8845**	**.3281**	.5812	.3385	**.3331**	**.4809**
Reconstruction	Original	-	-	-	.8125	.9985	.4795	.5978	.6245	.3910	.6506
	DAC	75	8	6	.7743	.9968	.4751	.5880	.6285	.4087	.6452
	Encodec	75	8	6	.7778	.9841	.4762	.5874	.5945	.3798	.6333
	SpeechTokenizer	50	8	4	.7812	.9940	.4582	.5790	.5495	.3609	.6204
	FACodec	80	6	4.8	.7083	.9856	.4766	.5933	.5615	.3632	.6148
	Encodec	75	2	1.5	.6667	.9516	**.4671**	.5852	.5715	.3656	.6013
	DAC	75	2	1.5	.6667	.9830	.4641	.5742	**.5955**	.3418	.6042
	SpeechTokenizer	50	3	1.5	.5729	.9771	.4364	.5414	.5150	.3670	.5683
	WavTokenizer	75	1	0.9	.6875	.9906	.4593	.5948	.5810	.3398	.6088
	SACodec(Ours)	75	2	1.5	**.7569**	**.9933**	.4563	**.5980**	.5825	**.3997**	**.6311**

Table 2: Semantic representation evaluation on the ARCH benchmark. We report classification accuracy for both the **Compressed Domain** and the **Reconstruction Domain**. Note that SACodec, WavTokenizer, and SpeechTokenizer were trained only on speech data, whereas DAC and Encodec utilized multi-domain (speech, music, audio) training data.

Model Configuration	**Reconstruction Quality**				**Semantic Acc. (C/R)**
	UTMOS ↑	PESQ ↑	STOI ↑	F1 ↑	$Avg.$ ↑
SACodec (Full Model, K_1=1000, K_2=1024)	4.0373	2.6937	**.9317**	.9381	**.4809** / .6311
Ablation on Semantic Anchor ($\mathbf{Q}_1$)					
w/o Q1 (SimVQ-only)	4.0132	2.6614	.9301	.9369	.3494 / .6198
w/ Random Learnable Codebook [a]	3.9051	2.5967	.9254	.9337	.3356 / .5966
w/ Smaller Anchor (K_1=500) [b]	3.8610	2.4981	.9208	.9307	.4786 / .5823
Ablation on Residual Activator ($\mathbf{Q}_2$)					
w/o Q2 (Semantic-only)	3.9461	2.3504	.9058	**.9407**	.5065 / .5765
w/ Larger Residual (K_2=2048)	**4.0402**	**2.7010**	.9308	.9379	.4789 / **.6402**

[a] Fixed mHuBERT codebook replaced with a standard learnable VQ.
[b] Semantic anchor size (K_1) reduced to 500, using HuBERT-L9 k-means centroids. This size was chosen as it is the only other publicly available, comparable semantic codebook.

Table 3: Ablation study of SACodec's core components and design choices. We evaluate reconstruction quality on LibriTTS test-clean and report mean semantic accuracy on the ARCH benchmark in the **C**ompressed / **R**econstruction domains.

producing tokens with superior semantic expressiveness. This is made possible by our asymmetric dual-quantizer, a design that overcomes the limitations of traditional VQ by synergistically anchoring semantics and activating residuals. Ultimately, SACodec provides a blueprint for a new generation of codecs, demonstrating that fidelity, semantics, and architectural simplicity can be achieved in unison for modern Speech Language Models.

Acknowledgments

This work was supported by the Beijing Xiaomi Mobile Software Co., Ltd, Beijing, China. In addition, the work leading to this research was supported by the National Natural Science Foundation of China under Grant No. U25A20447 and No. 62571184, the Science and Technology Innovation Program of Hunan Province under Grant No. 2025RC6003, the Guangdong Basic and Applied Basic Research Foundation under Grant No. 2024A1515010112, the Changsha Science and Tech-

nology Bureau Foundation under Grant No. kq2402082, and the Shenzhen Natural Science Foundation under Grant No. JCYJ20250604190534043.

References

Borsos, Z.; Marinier, R.; Vincent, D.; Kharitonov, E.; Pietquin, O.; Sharifi, M.; Roblek, D.; Teboul, O.; Grangier, D.; Tagliasacchi, M.; et al. 2023. AudioLM: A Language Modeling Approach to Audio Generation. *IEEE/ACM transactions on audio, speech, and language processing*, 31: 2523–2533.

Chen, S.; Liu, S.; Zhou, L.; Liu, Y.; Tan, X.; Li, J.; Zhao, S.; Qian, Y.; and Wei, F. 2024. VALL-E 2: Neural Codec Language Models are Human Parity Zero-Shot Text to Speech Synthesizers. *arXiv preprint arXiv:2406.05370*.

Chen, S.; Wang, C.; Chen, Z.; Wu, Y.; Liu, S.; Chen, Z.; Li, J.; Kanda, N.; Yoshioka, T.; Xiao, X.; et al. 2022. WavLM: Large-Scale Self-Supervised Pre-Training for Full Stack Speech Processing. *IEEE Journal of Selected Topics in Signal Processing*, 16(6): 1505–1518.

Chen, S.; Wang, C.; Wu, Y.; Zhang, Z.; Zhou, L.; Liu, S.; Chen, Z.; Liu, Y.; Wang, H.; Li, J.; He, L.; Zhao, S.; and Wei, F. 2025. Neural Codec Language Models are Zero-Shot Text to Speech Synthesizers. *IEEE Transactions on Audio, Speech and Language Processing*, 33: 705–718.

Défossez, A.; Copet, J.; Synnaeve, G.; and Adi, Y. 2023. High Fidelity Neural Audio Compression. *Transactions on Machine Learning Research*. 15 pages.

Défossez, A.; Mazaré, L.; Orsini, M.; Royer, A.; Pérez, P.; Jégou, H.; Grave, E.; and Zeghidour, N. 2024. Moshi: a speech-text foundation model for real-time dialogue. *arXiv preprint arXiv:2410.00037*.

Ding, D.; Ju, Z.; Leng, Y.; Liu, S.; Liu, T.; Shang, Z.; Shen, K.; Song, W.; Tan, X.; Tang, H.; et al. 2025. Kimi-Audio Technical Report. *arXiv preprint arXiv:2504.18425*.

Ding, X.; Zhang, X.; Ma, N.; Han, J.; Ding, G.; and Sun, J. 2021. RepVGG: Making VGG-Style ConvNets Great Again. In *Proc. 34th IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 13733–13742. Virtual.

Du, Z.; Gao, C.; Wang, Y.; Yu, F.; Zhao, T.; Wang, H.; Lv, X.; Wang, H.; Ni, C.; Shi, X.; et al. 2025. CosyVoice 3: Towards In-the-wild Speech Generation via Scaling-up and Post-training. *arXiv preprint arXiv:2505.17589*.

Esser, P.; Rombach, R.; and Ommer, B. 2021. Taming Transformers for High-Resolution Image Synthesis. In *Proc. 34th IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 12873–12883. Virtual.

Goodfellow, I. J.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A. C.; and Bengio, Y. 2014. Generative Adversarial Nets. In *Proc. 28th Conference on Neural Information Processing Systems (NeurIPS)*, 2672–2680. Montreal, Quebec, Canada.

Guo, Y.; Li, Z.; Du, C.; Wang, H.; Chen, X.; and Yu, K. 2024. LSCodec: Low-Bitrate and Speaker-Decoupled Discrete Speech Codec. *arXiv preprint arXiv:2410.15764*.

Guo, Y.; Li, Z.; Wang, H.; Li, B.; Shao, C.; Zhang, H.; Du, C.; Chen, X.; Liu, S.; and Yu, K. 2025. Recent Advances in Discrete Speech Tokens: A Review. *arXiv preprint arXiv:2502.06490*.

Hochreiter, S.; and Schmidhuber, J. 1997. Long Short-Term Memory. *Neural computation*, 9(8): 1735–1780.

Hsu, W.-N.; Bolte, B.; Tsai, Y.-H. H.; Lakhotia, K.; Salakhutdinov, R.; and Mohamed, A. 2021. HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units. *IEEE/ACM transactions on audio, speech, and language processing*, 29: 3451–3460.

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.; and Zhao, Z. 2025. WavTokenizer: an Efficient Acoustic Discrete Codec Tokenizer for Audio Language Modeling. In *Proc. 13th International Conference on Learning Representations (ICLR)*. Singapore. 15 pages.

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 *Proc. 63th Conference of the Association for Computational Linguistics (ACL)*, 19112–19124. Vienna, Austria.

Ju, Z.; Wang, Y.; Shen, K.; Tan, X.; Xin, D.; Yang, D.; Liu, Y.; Leng, Y.; Song, K.; Tang, S.; et al. 2024. NaturalSpeech 3: Zero-Shot Speech Synthesis with Factorized Codec and Diffusion Models. In *Proc. 41th International Conference on Machine Learning (ICML)*, 25697–25705. Vienna, Austria.

Kharitonov, E.; Vincent, D.; Borsos, Z.; Marinier, R.; Girgin, S.; Pietquin, O.; Sharifi, M.; Tagliasacchi, M.; and Zeghidour, N. 2023. Speak, Read and Prompt: High-Fidelity Text-to-Speech with Minimal Supervision. *Transactions of the Association for Computational Linguistics*, 11: 1703–1718.

Kong, J.; Kim, J.; and Bae, J. 2020. HiFi-GAN: Generative Adversarial Networks for Efficient and High Fidelity Speech Synthesis. In *Proc. 34th Conference on Neural Information Processing Systems (NeurIPS)*, 17022–17033. Virtual.

Kumar, R.; Seetharaman, P.; Luebs, A.; Kumar, I.; and Kumar, K. 2023. High-Fidelity Audio Compression with Improved RVQGAN. In *Proc. 37th Annual Conference on Neural Information Processing Systems (NeurIPS)*, 27980–27993. New Orleans, LA, USA.

Lee, A.; Gong, H.; Duquenne, P.-A.; Schwenk, H.; Chen, P.-J.; Wang, C.; Popuri, S.; Adi, Y.; Pino, J. M.; Gu, J.; and Hsu, W.-N. 2022. Textless Speech-to-Speech Translation on Real Data. In *Proc. 19th Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, 860–872. Seattle, WA, United States.

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 *Proc. 25th Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 3390–3394. Kos, Greece.

Liu, H.; Xu, X.; Yuan, Y.; Wu, M.; Wang, W.; and Plumbley, M. D. 2024. SemantiCodec: An Ultra Low Bitrate Semantic

Audio Codec for General Sound. *IEEE Journal of Selected Topics in Signal Processing*, 18(8): 1448–1461.

Ma, Z.; Song, Y.; Du, C.; Cong, J.; Chen, Z.; Wang, Y.; Wang, Y.; and Chen, X. 2025. Language Model Can Listen While Speaking. In *Proc. 39th AAAI Conference on Artificial Intelligence (AAAI)*, 24831–24839. Philadelphia, PA, USA.

Mousavi, P.; Maimon, G.; Moumen, A.; Petermann, D.; Shi, J.; Wu, H.; Yang, H.; Kuznetsova, A.; Ploujnikov, A.; Marxer, R.; et al. 2025. Discrete Audio Tokens: More Than a Survey! *arXiv preprint arXiv:2506.10274*.

Pratap, V.; Xu, Q.; Sriram, A.; Synnaeve, G.; and Collobert, R. 2020. MLS: A Large-Scale Multilingual Dataset for Speech Research. In Meng, H.; Xu, B.; and Zheng, T. F., eds., *Proc. 21th Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 2757–2761. Shanghai, China.

Quatra, M. L.; Koudounas, A.; Vaiani, L.; Baralis, E.; Cagliero, L.; Garza, P.; and Siniscalchi, S. M. 2024. Benchmarking Representations for Speech, Music, and Acoustic Events. In *Proc. 49th IEEE International Conference on Acoustics, Speech, and Signal Processing Workshops (ICASSPW)*, 505–509. Seoul, Republic of Korea.

Series, B. 2014. Method for the subjective assessment of intermediate quality level of audio systems. *International Telecommunication Union Radiocommunication Assembly*, 2. 36 pages.

Siuzdak, H. 2024. Vocos: Closing the Gap between Time-Domain and Fourier-based Neural Vocoders for High-quality Audio Synthesis. In *Proc. 12th International Conference on Learning Representations (ICLR)*. Vienna, Austria. 13 pages.

Xin, D.; Tan, X.; Takamichi, S.; and Saruwatari, H. 2024. BigCodec: Pushing the Limits of Low-Bitrate Neural Speech Codec. *arXiv preprint arXiv:2409.05377*.

Xu, J.; Guo, Z.; He, J.; Hu, H.; He, T.; Bai, S.; Chen, K.; Wang, J.; Fan, Y.; Dang, K.; et al. 2025. Qwen2.5-Omni Technical Report. *arXiv preprint arXiv:2503.20215*.

Yang, D.; Guo, H.; Wang, Y.; Huang, R.; Li, X.; Tan, X.; Wu, X.; and Meng, H. 2024. UniAudio 1.5: Large Language Model-Driven Audio Codec is A Few-Shot Audio Task Learner. In *Proc. 38th Conference on Neural Information Processing Systems (NeurIPS)*. Vancouver, BC, Canada. 13 pages.

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 preprint arXiv:2305.02765*.

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.

Zen, H.; Dang, V.; Clark, R.; Zhang, Y.; Weiss, R. J.; Jia, Y.; Chen, Z.; and Wu, Y. 2019. LibriTTS: A Corpus Derived from LibriSpeech for Text-to-Speech. *arXiv preprint arXiv:1904.02882*.

Zhang, D.; Li, S.; Zhang, X.; Zhan, J.; Wang, P.; Zhou, Y.; and Qiu, X. 2023. SpeechGPT: Empowering Large Language Models with Intrinsic Cross-Modal Conversational Abilities. In *Proc. 61th Conference of the Association for Computational Linguistics: Findings (EMNLP)*, 15757–15773. Singapore.

Zhang, X.; Zhang, D.; Li, S.; Zhou, Y.; and Qiu, X. 2024a. SpeechTokenizer: Unified Speech Tokenizer for Speech Language Models. In *Proc. 12th International Conference on Learning Representations (ICLR)*. Vienna, Austria. 12 pages.

Zhang, Z.; Dong, Z.; Xu, W.; and Han, J. 2025a. Re-Parameterization of Lightweight Transformer for On-Device Speech Emotion Recognition. *IEEE Internet of Things Journal*, 12(4): 4169–4182.

Zhang, Z.; Wu, Y.; Dong, Z.; Xiang, W.; Shen, S.; and Schuller, B. W. 2025b. SSE: A Speaking Style Extractor Based on Fine-Grained Contrastive Learning between Speech and Descriptive Text. In *Proc. 50th IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 1–5. Hyderabad, India.

Zhang, Z.; Xu, W.; Dong, Z.; Wang, K.; Wu, Y.; et al. 2024b. ParaLBench: a Large-Scale Benchmark for Computational Paralinguistics over Acoustic Foundation Models. *IEEE Transactions on Affective Computing*, 1–17.

Zhu, L.; Wei, F.; Lu, Y.; and Chen, D. 2024a. Scaling the Codebook Size of VQGAN to 100,000 with a Utilization Rate of 99%. In *Proc. 38th Annual Conference on Neural Information Processing Systems (NeurIPS)*, 12612–12635. Vancouver, BC, Canada.

Zhu, Y.; Li, B.; Xin, Y.; and Xu, L. 2024b. Addressing Representation Collapse in Vector Quantized Models with One Linear Layer. *arXiv preprint arXiv:2411.02038*.