Cross-Modal Aligned Streaming Single-Codebook Speech Codec

Anonymous ACL submission

Abstract

Discrete speech representations are critical for modern generative speech tasks and cross-modal modeling. However, current neural codecs often produce tokens that are either semantically redundant—entangled with paralinguistic variations like timbre—or structured in complex multi-codebook hierarchies that increase the complexity of downstream modeling. To bridge this gap, we propose **SecoustiCodec**, a streaming speech codec designed to extract *disentangled*, *single-codebook* discrete representations via cross-modal alignment. Unlike prior works relying on distillation from acoustic models, we introduce a frame-level text-speech contrastive learning framework that strictly aligns acoustic frames with linguistic units, effectively purging paralinguistic variance from the semantic codebook. To maintain high-fidelity reconstruction without compromising semantic purity, we explicitly model global paralinguistic attributes to complement the semantic tokens, allowing the decoder to synthesize fine-grained acoustic details from disentangled representations. Furthermore, we propose a semantic-only quantization mechanism combining Variational Autoencoders (VAE) and Finite Scalar Quantization (FSQ) to maximize codebook utilization and mitigate the long-tail distribution issue. SecoustiCodec supports low-latency streaming and achieves state-of-the-art reconstruction quality (PESQ 1.77/2.58 at 0.27/1 kbps). Audio samples are available at https://anonymous.4open.science/w/SecoustiCodec_Page-86F2. Code and models are provided at https://anonymous.4open.science/r/SecoustiCodec-BE3E.

1 Introduction

The intersection of speech processing and natural language processing has garnered significant attention, driven by the need for unified modeling of speech and text in generative tasks such as Text-to-Speech (TTS) (Wang et al., 2023; Zhang et al., 2023b), Automatic Speech Recognition (ASR) (Bai et al., 2024), and spoken dialogue systems (Défossez et al., 2024; Zeng et al., 2024). Central to this convergence is the neural speech codec, which discretizes continuous speech waveforms into a sequence of compact tokens. These discrete representations function analogously to text tokens, enabling the application of advanced sequence modeling techniques (e.g., Transformers) to the speech domain.

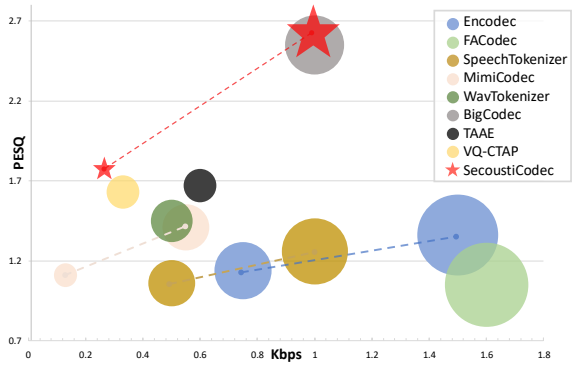


Figure 1: Comparison of different speech codecs operating below 2 kbps. The y-axis represents reconstruction quality (PESQ), while the x-axis indicates compression level (kbps). Circle sizes correspond to the number of discrete tokens encoded per second. SecoustiCodec supports streaming and claims SOTA performance in low-bitrate. Although BigCodec achieves comparable results, it neither supports causal streaming nor maintains parameter efficiency comparable to SecoustiCodec.

Despite their critical role, existing codecs face a fundamental dilemma between reconstruction fidelity and semantic disentanglement, which complicates downstream generative modeling. Traditional acoustic codecs like SoundStream (Zeghidour et al., 2021) and EnCodec (Défossez et al., 2022) prioritize fidelity by using Residual Vector Quantization (RVQ) with multiple codebooks.

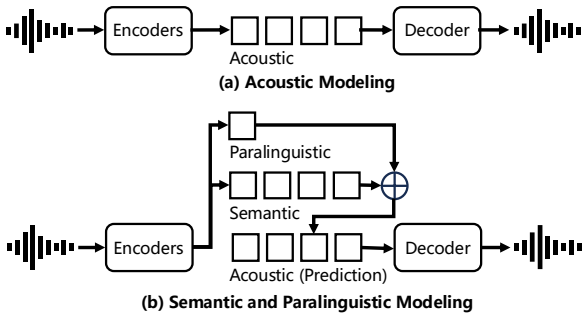


Figure 2: Conceptual overview of the acoustic-constrained disentanglement strategy. SecoustiCodec utilizes a learned continuous acoustic representation (A) to supervise the decomposition of speech into frame-level discrete semantic tokens (S) and a global paralinguistic embedding (G). By explicitly modeling paralinguistic attributes (e.g., timbre) as a **complementary** stream, we relieve the semantic codebook of acoustic variance. This design enforces robust semantic decoupling while ensuring that the combined representations contain sufficient information for high-fidelity acoustic reconstruction.

However, predicting multiple hierarchical tokens per frame significantly increases the sequence length and computational complexity for downstream autoregressive models (Wang et al., 2023). Conversely, semantic-centric codecs aim to extract a single sequence of semantic tokens. Recent approaches like SpeechTokenizer (Zhang et al., 2023a) and MimiCodec (Défossez et al., 2024) attempt to distill semantic information from pre-trained self-supervised models (e.g., HuBERT (Hsu et al., 2021), WavLM (Chen et al., 2022)). While effective, these pre-trained representations are inherently entangled with paralinguistic information (e.g., speaker identity, emotion), resulting in "semantic" tokens that still carry excessive acoustic variance. This entanglement forces generative models to learn redundant acoustic details, reducing their efficiency and controllability.

To address these limitations, we argue that an ideal discrete speech representation should possess three properties: (1) Single-Codebook Compactness: reducing sequence length for efficient modeling; (2) True Semantic Disentanglement: decoupling linguistic content from speaker timbre; and (3) Streaming Capability: supporting real-time interaction. We present **SecoustiCodec**, a cross-modal aligned streaming codec that satisfies these criteria. Instead of relying on acoustic distillation, we leverage text-speech contrastive learning to explicitly align speech frames with phoneme sequences. This cross-modal supervision forces the encoder to discard paralinguistic variations that do not correlate with text, yielding pure semantic tokens. Crucially, to ensure high-fidelity reconstruction from these semantic tokens, we introduce a complementary global paralinguistic encoder. We posit that *Semantic Content* and *Paralinguistic Style* are complementary: by explicitly feeding speaker attributes (e.g., timbre) to the decoder, we relieve the semantic codebook of the burden of carrying acoustic details. This allows for a single-codebook design that achieves high reconstruction quality previously possible only with multi-codebook methods.

Our main contributions are summarized as follows:

- We propose a cross-modal disentanglement framework that utilizes frame-level contrastive learning to align speech with text. This approach effectively purges paralinguistic information from the semantic codebook, facilitating efficient sequence modeling.
- We introduce a complementary modeling paradigm where a global paralinguistic embedding works in tandem with semantic tokens. This architecture bridges the information gap between semantic inputs and acoustic outputs, enabling high-fidelity reconstruction (SOTA PESQ of 2.58 at 1 kbps) with a single codebook.
- We design a VAE-FSQ hybrid quantization method that alleviates the long-tail distribution problem of discrete tokens, achieving 98.06% codebook utilization. Combined with an acoustic-constrained multi-stage optimization strategy, our model ensures robust convergence.
- We demonstrate that SecoustiCodec supports low-latency streaming (causal architecture) while outperforming existing baselines in both objective metrics and subjective listening tests.

We discuss related works in detail in Appendix A.

2 Method

2.1 Architecture Framework

As illustrated in Figure 3, SecoustiCodec extracts disentangled representations from speech (S_{in}) and phonemes (P_{in}) through three causal streams:

Feature Extraction: (1) Acoustic (A): A continuous representation extracted by a convolutional encoder: $A = \text{Proj}(\text{SpeechEncoder}(S_{in}))$. (2) Phoneme ($P$): Extracted from text and aligned via

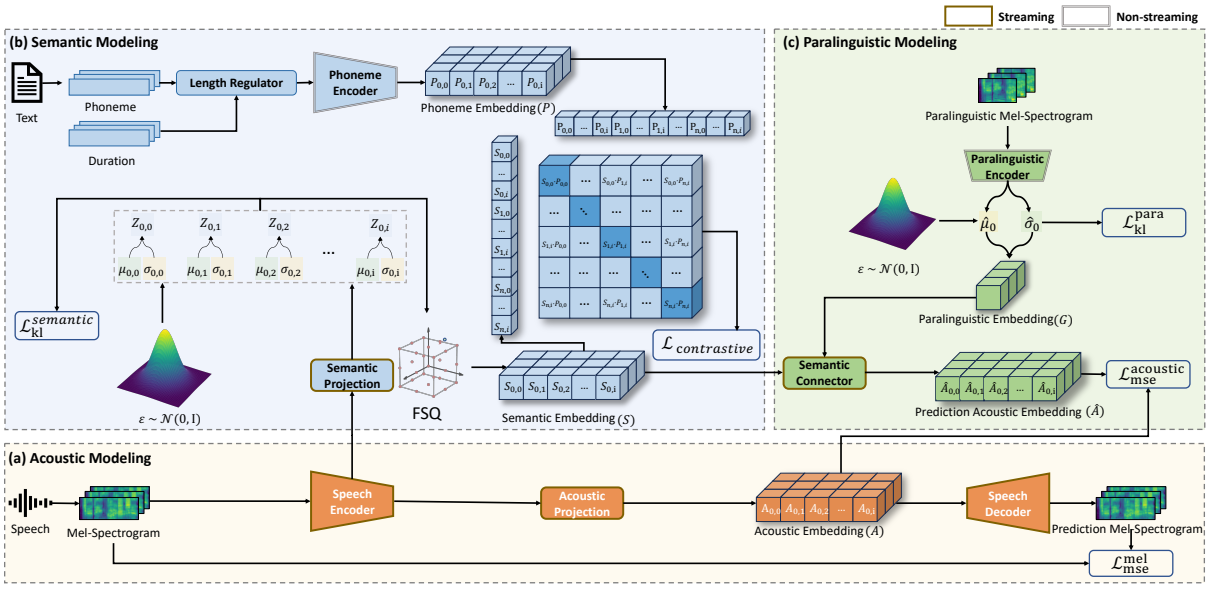


Figure 3: SecoustiCodec includes three modeling processes: (a) *Acoustic Modeling*, (b) *Semantic Modeling* and (c) *Paralinguistic Modeling*. Modules outlined in red operate in a streaming manner, while those in blue are non-streaming. Phoneme embeddings (P) are extracted from text, and target semantic embeddings (S), acoustic embeddings (A), and paralinguistic embeddings (G) are extracted from speech. (P) and (S) are used to construct token-acoustic contrastive loss, which learns frame-level (dis)similarity between a batch of speech and text pairs. In the inference process, Acoustic Projection is not required; instead, semantic embedding and paralinguistic embedding are used to predict acoustic embedding. The mean values (μ and $\hat{\mu}$) from the VAE structure are directly used as inputs during inference, bypassing stochastic sampling. The name "SecoustiCodec" signifies a **codec** that supports both **se**mantic and a**coustic** encoding.

a length regulator: $P = \text{PhEncoder}(P_{in})$. (3) Paralinguistic (G): A global embedding captured from a 3s window (G_{in}) using a VAE encoder. To prevent KL collapse, we apply a hinged KL loss:

$$\mathcal{L}_{kl}^{para} = \max(0, D_{KL}[\mathcal{N}(\hat{\mu}, \hat{\sigma}^2)||\mathcal{N}(0, I)] - \Delta) \quad (1)$$

Reconstruction The decoder reconstructs the mel-spectrogram from acoustic features: $S_p = \text{SpeechDecoder}(A)$. To enforce semantic completeness, a Semantic Connector predicts the acoustic embedding A from the discrete semantic tokens S (defined in Sec. 2.2) and style G: $\hat{A} = \text{Connector}(S, G)$. We minimize the reconstruction losses: $\mathcal{L}_{recon} = ||\hat{A} - A||_2^2 + ||S_p - S_{in}||_2^2$.

2.2 VAE-FSQ Semantic Quantization

To obtain the semantic code S, we employ a hybrid **VAE-FSQ** approach. The speech encoder output is projected to a latent distribution $\mathcal{N}(\mu, \sigma)$, from which we sample a frame-level vector z using the reparameterization trick. z is then quantized via Finite Scalar Quantization (FSQ) to alleviate codebook collapse:

$$S = \text{Proj}_{\text{up}}(\text{round}(\lfloor L/2 \rfloor \tanh(\text{Proj}_{\text{down}}(z)))) \quad (2)$$

This projects z into a low-rank space with L levels, implicitly forming a codebook of size L^d. A KL regularization $\mathcal{L}_{kl}^{semantic}$ is applied similarly to Eq. (1).

2.3 Cross-Modal Token-Acoustic Contrastive Learning

As shown in Figure 3 (b), to extract frame-level representations, S and P within a batch are reshaped into 2D matrices S_{re} and P_{re}, where S_{re} and $P_{re} \in \mathbb{R}^{(B*T_s)\times d}$. This approach is beneficial for contrastive learning, as it increases the number of sample pairs per step, thereby enhancing the learning process.

With S_{re} and P_{re} now comparable, we can measure their similarity: $C = \tau * (S_{re} \cdot P_{re}^{\top})$ where τ is a temperature parameter used to scale the range of logits. The resulting similarity matrix $C \in \mathbb{R}^{(B*T_s)\times(B*T_s)}$ contains $(B * T_s)$ correct pairs along the diagonal and $(B * T_s)^2 - (B * T_s)$ incorrect pairs in the off-diagonal elements. As the extracted intermediate representation includes contextual information, only the current frame corresponds to a positive sample. The contrastive loss is calculated:

$$\mathcal{L}_{contrastive} = 0.5 * (\ell_{speech}(C) + \ell_{phoneme}(C)) \quad (3)$$

where $\ell = \frac{1}{B*T_s} \sum_{i=0}^{B*T_s} \log diag(softmax(C))$ along the speech and phoneme axes, respectively. This symmetric cross-entropy loss ($\mathcal{L}_{contrastive}$) is computed over the similarity matrix to jointly train the speech encoder and the phoneme encoder, enabling the model to learn meaningful representations from both modalities simultaneously. Note that the phoneme encoder and length regulator are employed exclusively during training to enforce alignment constraints. During inference, only the speech encoder is required to extract semantic tokens.

2.4 Multi-Stage Optimization Strategy

To ensure stable convergence, we introduce an acoustic-constrained multi-stage optimization strategy. The detailed procedure is outlined in Algorithm 1 in Appendix C.

Stage 1 (Acoustic Modeling): We first pre-train the speech encoder and decoder exclusively on a mel-spectrogram reconstruction task ($\mathcal{L}_{mse}^{mel}$). This initial stage establishes a robust acoustic representation that serves as a fixed target for the subsequent stage.

Stage 2 (Disentanglement Modeling): With the acoustic modules frozen, we then train the semantic and paralinguistic components. The primary objective is to make the combined semantic and paralinguistic representations predict the pre-trained acoustic representation A, guided by an acoustic MSE loss ($\mathcal{L}_{mse}^{acoustic}$) and the cross-modal contrastive loss ($\mathcal{L}_{contrastive}$). To prevent posterior collapse, the KL divergence terms for both the semantic ($\mathcal{L}_{kl}^{semantic}$) and paralinguistic ($\mathcal{L}_{kl}^{para}$) VAEs are progressively introduced via linear warm-up schedules.

3 Experiments Procedures

3.1 Model Details

Our architecture draws inspiration from MimiCodec(Défossez et al., 2024). All components employ causal operations to ensure streaming encoding and decoding. The speech encoder/decoder is built using causal SeaNet encoder blocks (Tagliasacchi et al., 2020), the speech encoder consists of 80-channel convolutional layers with kernel size 7, ELU activation, and dilation base 2. The speech decoder mirrors this architecture with transposed convolutions for mel-spectrogram reconstruction. Both use 1 residual layer with true skip connections and a compression factor of 2. The acoustic projection employs a causally masked transformer with 8 layers and 8-head attention, featuring RoPE positional encoding (Su et al., 2024). The 512-dimensional model includes layer normalization and a 2048-unit feedforward network, processing 250-frame context windows with convolutional layout integration. The semantic projection combines a causal transformer (identical architecture to acoustic projection) with variational inference. Two linear projections (μ and σ) map the transformer's 512D output to latent variables using the reparameterization trick. The semantic connector shares architectural parameters with the acoustic projection module, forming an 8-layer transformer with identical attention mechanisms and normalization schemes to maintain dimensional consistency across semantic representations. The phoneme encoder employs a convolutional layer (ReLU-activated) followed by 4 transformer layers and linear projection. The paralinguistic encoder utilizes a VAE structure with 6 convolutional layers and SE-ResNet blocks (Hu et al., 2018). All encoder outputs undergo layer normalization before fusion. The vocoder used in this experiment is HifiGAN (Kong et al., 2020). For the multi-stage optimization strategy, the parameters are as follows: $stage1_{end} = 1e4$, $kl_{start}^{semantic} = 2e4$, $kl_{start}^{para} = 2e4$, $kl_{end}^{semantic} = 3e4$, $kl_{end}^{para} = 3e4$, $kl_{upper}^{semantic} = 1e-5$, $kl_{upper}^{para} = 1e-5$, $\alpha = 1$, $\beta = 1e-5$. The model is trained using 8 NVIDIA TESLA A800 80GB GPUs, with a batch size of 64 per GPU. Adam(Kingma and Ba, 2014) is used as the optimizer, with an initial learning rate of 2e-4. The number of sample pairs per step ranges from 4,000 to 32,000 during the training process.

3.2 Compared Method and Tasks

To evaluate the performance of the proposed model, we conduct experiments on speech reconstruction task. We compare **SecoustiCodec** with eight representative Codec models, as summarized in Table 1. The table lists key attributes of each model, including bitrate, frame rate, number of quantizers, tokens per second, parameter size, training data duration, and whether the model employs a causal structure supporting streaming encoding and decoding. The comparison includes classic streaming Codec

models like **Encodec**[1], semantic disentanglement models like **SpeechTokenizer**[2] and **MimiCodec** (Moshi)[3], as well as **FACodec** (NaturalSpeech 3)[4], which disentangles attributes through supervised tasks, and single-codebook models like **VQ-CTAP**, **WavTokenizer**[5], **BigCodec**[6], and **TAAE**[7].

We also compare multi-codebook models under different bitrate conditions to demonstrate the effectiveness of SecoustiCodec. Despite SecoustiCodec's denser structure, it does not increase the parameter count. Additionally, as shown in Table 2, we perform extensive ablation studies to analyze various factors. These include the quantization method, the use of a causal structure for streaming support, the inclusion of pitch features alongside Mel spectrograms, the application of a multi-stage optimization strategy, and the dimensionality of the acoustic embedding. Finally, we also evaluate the performance of SecoustiCodec in single-speaker scenarios and Voice Conversion (VC) task.

3.3 Datasets

For the labeled text-speech paired data, we integrate our internal dataset with the AISHELL-3 dataset (Shi et al., 2020) and the LibriTTS dataset (Zen et al., 2019), resulting in a combined total of 1,000 hours of recordings from 3,000 speakers. All speech waveforms are sampled at 22kHz and converted to 80-band mel spectrograms with a window size of 1024 and a hop size of 256.

3.4 Evaluation Metrics

We evaluate the model using several objective and subjective metrics to ensure comprehensive performance assessment. These metrics include: Wideband PESQ (Perceptual Evaluation of Speech Quality, downsampled to 16kHz)[8], MUSHRA (Multiple Stimuli with Hidden Reference and Anchor), Speaker Similarity[9], Emotion Similarity[10], LSD (Log-Spectral Distance), MCD (Mel-Cepstral Distortion), MSEP (Mean Squared Error of Pitch), MR (Voiced/unvoiced Mismatch Rate), NISQA (Speech Quality and Naturalness Assessment)[11], and WER (Word Error Rate)[12].

4 Results and Analysis

4.1 Structure Comparison

Table 1 compares the structure of SecoustiCodec with mainstream codecs. SecoustiCodec achieves a low bitrate of 0.27kbps, a frame rate of 20Hz and a token rate of 20 tokens/s, second only to MimiCodec. In terms of quantization methods, it adopts the proposed VAE-FSQ approach instead of the RVQ or VQ structures used by other models. This enables single-codebook semantic encoding while supporting continuous-value acoustic encoding. Although the model has a relatively high parameter count, it remains smaller than FACodec, SpeechTokenizer, BigCodec and TAAE. Regarding training data, SecoustiCodec uses 1,000 hours of data, similar to SpeechTokenizer, MimiCodec and BigCodec, but significantly less than FACodec's 50,000 hours and TAAE's 100,000 hours. Additionally, it employs a causal structure to support streaming encoding and decoding. Experimental results demonstrate that SecoustiCodec achieves optimal performance under conditions of causal structure, single-codebook encoding, low bitrate, and limited training data. In Table 1 , we present the initial latency and real-time factors (RTF) for all causal structure-supported streaming models. The real-time factor is defined as the ratio between the processing time and the speech duration, so that it is less than one when the method is faster than real time. We profiled all models on a single thread of a V800 GPU. SecoustiCodec exhibits an initial latency of 12.08 ms, which is lower than that of Encodec and MimiCodec. In terms of real-time factor, Encodec performs significantly better than SecoustiCodec. Notably, both Encodec and SecoustiCodec have an encoding RTF of 0.002. However, the decoding process for SecoustiCodec includes time-consuming vocoder operations, which increases its decoding real-time factor. Excluding the vocoder processing time, SecoustiCodec achieves a decoding real-time factor of 0.001, which is lower than that of Encodec.

4.2 Evaluation of Speech Reconstruction

Table 1 presents the results of the speech reconstruction task. To demonstrate the effectiveness

[1] https://github.com/facebookresearch/encodec
[2] https://github.com/ZhangXInFD/SpeechTokenizer
[3] https://huggingface.co/kyutai/mimi
[4] https://github.com/lifeiteng/naturalspeech3_facodec
[5] https://github.com/jishengpeng/WavTokenizer
[6] https://github.com/Aria-K-Alethia/BigCodec
[7] https://github.com/Stability-AI/stable-codec
[8] https://github.com/ludlows/PESQ
[9] https://github.com/resemble-ai/Resemblyzer
[10] https://huggingface.co/emotion2vec
[11] https://github.com/gabrielmittag/NISQA
[12] https://github.com/modelscope/FunASR

Table 1: Reconstruction Results of Different Codec Models including Subjective Evaluation

Model		GT	Encodec		FACodec		SpeechTokenizer		MimiCodec		VQ-CTAP	WavTokenizer	BigCodec	TAAE	SecoustiCodec	
Bitrate↓		\	1.5kbps	6kbps	1.6kbps	4.8kbps	1kbps	4kbps	0.55kbps	4.4kbps	0.33kbps	0.5kbps	1kbps	0.6kbps	0.27kbps	1kbps
Frame Rate↓		\	75Hz		80Hz		50Hz		12.5Hz		25Hz	40Hz	80Hz	25Hz	20Hz	80Hz
Nq↓		\	2	8	2	6	2	8	4	32	1	1	1	1	1	
Tokens/s↓		\	150	600	160	480	100	400	100	800	25	40	80	25	20	80
Param Size↓		\	57MB		406MB		396MB		303MB		105MB	308MB	608MB	3,636MB	379MB	
Training Dataset↓		\	17,000h		500,000h		1000h		1000h		11,000h	80,000h	1,000h	100,000h	1,000h	
Causal		\	✓		✗		✗		✓		✗	✗	✗	✗	✓	
Latency↓		\	13.39ms		\		\		84.4ms		\	\	\	\	12.08ms	
RTF↓	**Total**	\	0.004		\		\		0.056		\	\	\	\	0.04	
	Enc.	\	0.002		\		\		0.033		\	\	\	\	0.002	
	Dec.	\	0.002		\		\		0.023		\	\	\	\	0.038	
MUSHRA↑		99.00	38.25	61.50	28.50	75.20	32.40	63.80	39.50	81.00	44.80	41.50	70.50	46.20	49.60	**72.10**
PESQ↑		4.64	1.36	2.28	1.05	2.90	1.26	2.36	1.41	3.28	1.63	1.45	2.55	1.67	1.77	**2.58**
Spk Sim↑		1.00	0.9	0.98	0.80	0.98	0.74	0.97	0.85	0.97	0.89	0.86	**0.97**	0.83	0.92	0.95
Emo Sim↑		1.00	0.86	0.96	0.79	0.97	0.84	0.97	0.87	0.97	0.92	0.91	0.94	0.92	0.93	**0.97**
LSD↓		0.00	1.06	0.94	1.26	0.83	1.14	1.00	1.19	0.96	0.90	0.97	0.84	1.32	0.85	**0.77**
MCD↓		0.00	1.63	0.85	4.07	0.86	2.24	1.07	1.79	0.58	1.86	1.94	**0.94**	2.59	1.72	1.32
MSEP↓		0.00	36.03	17.06	630.71	12.15	40.83	18.93	43.28	9.10	29.00	39.29	18.26	28.57	39.66	**18.25**
Mismatchrate↓		0.00	0.18	0.13	0.28	0.05	0.11	0.06	0.09	0.05	0.07	0.08	0.06	0.07	0.07	**0.05**
WER↓		3.05	5.60	3.37	15.61	3.47	16.97	3.62	10.69	3.23	17.06	19.70	4.10	30.24	11.58	**3.99**

GT: **G**round **T**ruth Waveforms; Nq: **N**umber of **Q**uantizers; Causal: Streaming Support; RTF: **R**eal **T**ime **F**actor; ✓: Enabled; ✗: Disabled.
Best results from models with a single quantizer (hence directly comparable to SecoustiCodec) are in **bold**.

Table 2: Ablation Studies of SecoustiCodec

Model	Quant	Causal	F0	Stage	Acous-Dim	PESQ↑	Spk Sim↑	Emo Sim↑	LSD↓	MCD↓	MSEP↓	MR↓	NISQA↑	WER↓
Secousti-VQ	VQ-VAE	✓	✗	✓	256	1.28	0.87	0.92	0.96	2.31	54.54	0.09	2.72	47.28
Secousti-VQ w/o Stage	VQ-VAE	✓	✗	✗	256	1.20	0.87	0.91	0.98	2.56	71.89	0.11	2.66	56.28
Secousti-VQ w/o Causal	VQ-VAE	✗	✗	✓	256	1.44	0.62	0.8	1.69	5.42	46.47	0.18	2.51	14.31
Secousti-VQ w/o Causal w/o Stage	VQ-VAE	✗	✗	✗	256	1.10	0.74	0.84	1.03	3.19	155.64	0.15	1.86	66.77
Secousti-VQ w/ F0	VQ-VAE	✓	✓	✓	256	1.02	0.54	0.71	1.30	4.27	1286.17	0.52	0.97	98.38
Secousti-VQ w/ F0 w/o Stage	VQ-VAE	✓	✓	✗	256	1.03	0.58	0.67	1.50	4.28	519.01	0.44	1.35	100.00
Secousti-SimVQ	SimVQ	✓	✗	✓	256	1.65	0.90	0.92	0.89	1.84	**28.63**	**0.07**	3.07	16.95
Secousti-FSQ-8dim	FSQ	✓	✗	✓	8	1.53	0.93	0.94	0.89	1.99	48.53	**0.07**	**4.07**	14.00
Secousti-FSQ-64dim	FSQ	✓	✗	✓	64	1.57	0.90	0.91	0.88	1.89	54.18	0.09	3.23	18.67
Secousti-FSQ-256dim	FSQ	✓	✗	✓	256	1.71	0.90	0.92	0.88	1.78	30.16	**0.07**	3.13	15.13
Secousti-FSQ-64dim w/o Stage	FSQ	✓	✗	✗	64	1.65	0.91	0.92	0.87	1.82	40.18	0.08	3.18	13.23
Secousti-VAE-FSQ-64dim	VAE+FSQ	✓	✗	✓	64	1.60	0.90	0.92	0.87	1.86	49.78	0.08	3.23	17.04
SecoustiCodec(0.27kbps)	VAE+FSQ	✓	✗	✓	256	**1.77**	**0.92**	**0.93**	**0.85**	**1.72**	39.66	**0.07**	3.50	**11.58**

Quant: Quantization Method; **Causal**: Streaming Support; **F0**: Pitch Feature; **Stage**: Multi-stage Training; **Acous-Dim**: Acoustic Embedding Dimension.
✓: Enabled, ✗: Disabled. Bold row indicates our standard configuration.

of SecoustiCodec, we compare its performance with other single-codebook methods and multi-codebook methods under different bitrates.

SecoustiCodec (1kbps) achieves the best performance across multiple metrics, including MUSHRA, PESQ, EmoSim, LSD, F0, MR, and WER, surpassing some models even in scenarios where they use multiple codebooks. Crucially, the subjective MUSHRA results align remarkably well with the objective Wideband PESQ improvements. SecoustiCodec (1kbps) achieves a MUSHRA score of **72.10**, demonstrating high-fidelity reconstruction comparable to state-of-the-art models like BigCodec. SecoustiCodec (0.27kbps 20Hz) is the model with the lowest bitrate and frame rate among single-codebook models. It achieves a MUSHRA score of **49.60**, significantly outperforming other low-bitrate candidates such as WavTokenizer (41.50) and VQ-CTAP (44.80). Despite the extremely low bitrate, SecoustiCodec (0.27kbps 20tokens/s) also outperforms multi-codebook models such as Encodec (1.5kbps 150tokens/s), FACodec (1.6kbps 160tokens/s), SpeechTokenizer (1kbps 100tokens/s), and MimiCodec (0.55kbps 100tokens/s) in terms of subjective quality.

Methods such as FACodec, SpeechTokenizer, and MimiCodec supervise the semantic encoding only in the first-dimensional VQ space, often incorporating ASR tasks or distillation tasks from Hubert(Hsu et al., 2021)/ WavLM(Chen et al., 2022). However, the representations from Hubert/ WavLM are not purely semantic and contain significant residual paralinguistic information. These models combine semantic encodings with other encodings during the decoding stage to reconstruct speech. This approach limits the disentanglement and reconstruction capability of the semantic encoding itself. During training, the model relies on the full set of encodings to extract acoustic information, which prevents the semantic encoding from independently ensuring semantic completeness. As a result, substantial paralinguistic information often

Table 3: Results of Single Speaker

Model	Bitrate	PESQ↑	Spk Sim↑	Emo Sim↑	LSD↓	MCD↓	F0↓	MR↓	WER↓
SecoustiCodec	0.27kbps	2.50	0.96	0.80	0.76	1.18	20.88	0.06	6.51
ft Single Speaker	1kbps	3.51	0.97	0.97	0.71	0.89	8.99	0.04	3.37

remains in the semantic encoding.

In contrast, SecoustiCodec benefits from its disentanglement of semantic encoding and its explicit modeling of paralinguistic information. This enables it to outperform other methods significantly in both SpkSim and EmoSim. During training, SecoustiCodec independently models semantic, acoustic, and paralinguistic information. Additionally, it incorporates a task where the semantic encoding predicts the acoustic encoding, using paralinguistic encoding as input to complement the semantic information. This approach helps the model learn pure semantic and paralinguistic representations. Consequently, SecoustiCodec achieves superior results in both subjective MUSHRA scores and objective reconstruction metrics such as PESQ, EmoSim, LSD, F0, MR, and WER. It is important to note that these results are obtained under challenging conditions, including the use of a causal structure, single-codebook encoding, high compression rates, and limited training data.

4.3 Ablation Studies

As shown in Table 2, we conduct extensive ablation experiments to validate various aspects of our model. Regarding quantization methods, SecoustiCodec outperforms other models, such as Secousti-SimVQ, Secousti-VQ, and Secousti-FSQ, across all metrics except MSEP, demonstrating the superiority of the VAE-FSQ quantization method. We also compare the codebook utilization of different quantization methods, with detailed distribution plots provided in Figure 5 of Appendix C. As illustrated, VAE-FSQ achieves the highest codebook utilization (98.06%) with a uniform distribution, significantly minimizing the long-tail effect compared to VQ-VAE and SimVQ. Additionally, VAE-FSQ minimizes the long-tail effect, with most token frequencies below 0.2%, which aids language model training in TTS and voice dialogue tasks. In terms of causal structure, we observe that removing the causal constraint (Secousti-VQ w/o Causal) yields superior performance in key metrics such as PESQ, MSEP, and WER compared to the standard Secousti-VQ. This performance gap represents an expected trade-off: non-causal models benefit significantly from bidirectional context (access to future information) for precise acoustic reconstruction and phoneme alignment. However, to satisfy the strict low-latency requirements of real-time dialogue systems, SecoustiCodec adopts a fully causal architecture. While this incurs a slight degradation in reconstruction quality compared to the non-causal upper bound, it is a necessary compromise to enable streaming capability. Regarding pitch features, adding pitch features to Secousti-VQ (Secousti-VQ w/ F0) negatively impacts nearly all metrics, particularly MSEP and WER, where it records the worst values among all ablation models. While pitch features aid acoustic modeling, they significantly interfere with semantic modeling capabilities. Concerning the multi-stage training strategy, models that do not employ this strategy, such as Secousti-VQ w/o Stage, Secousti-VQ w/o Causal w/o Stage, Secousti-VQ w/ F0 w/o Stage, and Secousti-FSQ-64dim w/o Stage, show significant degradation in all metrics. This is because introducing additional loss terms during the acoustic modeling stage affects the convergence of the $\mathcal{L}_{mse}^{mel}$, ultimately reducing the representational capacity of the generated acoustic embeddings. The presence of the $\mathcal{L}_{mse}^{acoustic}$ further diminishes the semantic encoding reconstruction ability. Regarding the dimensionality of acoustic embedding, we perform ablation experiments on the Secousti-FSQ structure. Since semantic encoding is mapped to an 8-dimensional embedding via a linear layer before FSQ, we aim for consistent dimensionality in acoustic encoding to reduce modeling complexity. However, we find that the dimensionality of the acoustic embedding significantly influences the model's overall capability. As evidenced by Secousti-FSQ-8dim, Secousti-FSQ-64dim, and Secousti-FSQ-256dim, increasing the acoustic embedding dimension improves all metrics. Ultimately, we select 256 as the acoustic encoding dimension for SecoustiCodec.

Table 4: Results of VC

Model	SpkSim↑	EmoSim↑	WER↓
SecoustiCodec-ASR-Loss	0.64	0.69	**10.51**
SecoustiCodec	**0.71**	**0.86**	11.58

4.4 Semantic-Paralinguistic Disentangle and Single-Speaker

We have added frame-level ablation experiments for ASR loss. For fairness, we replaced frame-level contrastive learning with frame-level phoneme classification loss (Secousti-ASR-Loss). To demonstrate the effect of semantic decoupling, we validated this on the VC task (replacing Paralinguistic Embedding at the decoder stage). As shown in Table 4, compared to ASR loss, frame-level contrastive loss achieved higher speaker similarity and emotional consistency, but Secousti-ASR performed better on the WER metric. Compared to methods that use ASR loss, frame-level contrastive learning demonstrates a stronger ability to incorporate cross-modal textual information for semantic disentanglement. Visualization of the synthesized spectrograms, F0 contours, and energy profiles can be found in Figure 4 in Appendix C. As shown in the visualization, the synthesized speech exhibits strong alignment with the target paralinguistic attributes across multiple frequency domains, demonstrating the effectiveness of our disentanglement strategy. This is attributed to the disentanglement of paralinguistic information in the semantic encoding and the paralinguistic modeling capability of the paralinguistic encoder in SecoustiCodec.

We conduct additional experiments in a single-speaker scenario. In tasks such as TTS and voice-based dialogue, it is often sufficient to use the timbre of a single target speaker. In practical applications, a decoder specifically trained for the target speaker is commonly used to enhance audio quality, speaker consistency, and expressiveness. As shown in Table 3, the fine-tuned single-speaker model achieves high performance across all metrics.

5 Analysis

We analyze the intrinsic properties of SecoustiCodec representations and their potential advantages over traditional multi-codebook approaches:

Sequence Compactness. Multi-codebook models (e.g., Encodec) typically rely on 8 parallel codebooks, which necessitates either hierarchical modeling or flattening strategies that drastically increase sequence length. By achieving high-fidelity reconstruction with a **single codebook**, SecoustiCodec offers an $8\times$ reduction in token count compared to flattened RVQ. Since Transformer complexity scales quadratically with sequence length, our compact representation provides a computationally efficient foundation for future autoregressive modeling, reducing the overhead of processing long speech contexts.

Disentanglement for Controllability. In standard codecs, semantic tokens often exhibit leakage of speaker-specific details (e.g., pitch), entangling content with timbre. While we leave downstream LLM training for future work, our analysis suggests that such entanglement forces models to learn redundant acoustic patterns. By explicitly purging paralinguistic variance via contrastive learning, SecoustiCodec produces speaker-agnostic semantic tokens. This structural disentanglement theoretically simplifies the modeling task for downstream systems, as they can focus solely on linguistic content generation while offloading timbre control to the global paralinguistic embedding.

Codebook Utilization Efficiency. High codebook utilization is a desirable property for discrete representations, ensuring maximum information capacity. As shown in Figure 5 (Appendix C), methods like VQ-VAE often suffer from "codebook collapse," where a large portion of the vocabulary remains unused. In contrast, our VAE-FSQ strategy achieves **98.06% utilization** with a near-uniform distribution. This high perplexity suggests that SecoustiCodec effectively exploits the discrete latent space, reducing representational redundancy.

6 Conclusion

We presented SecoustiCodec, a streaming speech codec that achieves state-of-the-art reconstruction fidelity at low bitrates (0.27/1 kbps) using a single codebook. By leveraging cross-modal contrastive learning and a novel VAE-FSQ quantization scheme, we successfully disentangled semantic content from paralinguistic attributes while maintaining high codebook utilization (98.06%). The introduction of explicit paralinguistic modeling ensures semantic completeness without compromising acoustic detail. Future work will focus on exploring unsupervised disentanglement to eliminate the dependency on labeled text-speech pairs and extending the model's adaptability to low-resource languages beyond English and Chinese.

7 Limitations

While SecoustiCodec demonstrates promising results, several limitations remain:

- **Dependency on Paired Data:** Unlike self-supervised methods (e.g., WavLM-based codecs) that train on raw audio, our semantic disentanglement relies on text-speech paired data for contrastive learning. This dependency limits the scale of training data compared to unsupervised approaches and constrains applicability to low-resource languages lacking high-quality transcriptions.
- **Generalization to Unseen Languages:** Our experiments are conducted primarily on English and Chinese datasets. The codebook's ability to represent phonemes or tonal patterns from unseen language families remains unverified, potentially leading to performance degradation in cross-lingual scenarios.
- **Trade-off in Causal Modeling:** To enable streaming capabilities, we adopt a strictly causal architecture. As observed in our ablation studies, this introduces a slight degradation in reconstruction quality compared to non-causal variants that utilize bidirectional context.

8 Ethical Considerations

The development of high-fidelity, disentangled speech codecs raises potential ethical concerns regarding misuse:

- **Voice Cloning and Deepfakes:** SecoustiCodec explicitly decouples timbre (paralinguistic) from content. While this benefits Voice Conversion (VC) and TTS, it also lowers the barrier for generating unauthorized voice clones or deepfakes. Malicious actors could potentially use the disentangled paralinguistic encoder to synthesize misleading audio resembling specific individuals.
- **Mitigation Strategies:** We strongly condemn the misuse of this technology. We advocate for the integration of invisible watermarking techniques in the decoding stage for future deployments.
- **Bias in Training Data:** The datasets used (LibriTTS, AISHELL-3) may contain demographic biases. The model might underperform for speakers with accents, dialects, or speech impediments not well-represented in the training corpora.

References

Ye Bai, Jingping Chen, Jitong Chen, Wei Chen, Zhuo Chen, Chuang Ding, Linhao Dong, Qianqian Dong, Yujiao Du, Kepan Gao, et al. 2024. Seed-asr: Understanding diverse speech and contexts with llm-based speech recognition. *arXiv preprint arXiv:2407.04675*.

Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, 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.

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

Alexandre Défossez, Laurent Mazaré, Manu Orsini, Amélie Royer, Patrick Pérez, Hervé Jégou, Edouard Grave, and Neil Zeghidour. 2024. Moshi: a speech-text foundation model for real-time dialogue. *arXiv preprint arXiv:2410.00037*.

Benjamin Elizalde, Soham Deshmukh, Mahmoud Al Ismail, and Huaming Wang. 2023. Clap learning audio concepts from natural language supervision. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Yaroslav Ganin and Victor Lempitsky. 2015. Unsupervised domain adaptation by backpropagation. In *International conference on machine learning*, pages 1180–1189. PMLR.

Cristina Gârbacea, Aäron van den Oord, Yazhe Li, Felicia SC Lim, Alejandro Luebs, Oriol Vinyals, and Thomas C Walters. 2019. Low bit-rate speech coding with vq-vae and a wavenet decoder. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 735–739. IEEE.

Andrey Guzhov, Federico Raue, Jörn Hees, and Andreas Dengel. 2022. Audioclip: Extending clip to image, text and audio. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 976–980. IEEE.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. 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.

Jie Hu, Li Shen, and Gang Sun. 2018. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7132–7141.

Shengpeng Ji, Ziyue Jiang, Wen Wang, Yifu Chen, Minghui Fang, Jialong Zuo, Qian Yang, Xize Cheng,

Zehan Wang, Ruiqi Li, et al. 2024. Wavtokenizer: an efficient acoustic discrete codec tokenizer for audio language modeling. *arXiv preprint arXiv:2408.16532*.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. 2021. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, pages 4904–4916. PMLR.

Zeqian Ju, Yuancheng Wang, Kai Shen, Xu Tan, Detai Xin, Dongchao Yang, Yanqing Liu, Yichong Leng, Kaitao Song, Siliang Tang, et al. 2024. Naturalspeech 3: Zero-shot speech synthesis with factorized codec and diffusion models. *arXiv preprint arXiv:2403.03100*.

Diederik P. Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980.

Jungil Kong, Jaehyeon Kim, and Jaekyoung Bae. 2020. Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis. *Advances in Neural Information Processing Systems*, 33:17022–17033.

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. *arXiv preprint arXiv:2405.00233*.

Julian D Parker, Anton Smirnov, Jordi Pons, CJ Carr, Zack Zukowski, Zach Evans, and Xubo Liu. 2024. Scaling transformers for low-bitrate high-quality speech coding. *arXiv preprint arXiv:2411.19842*.

Chunyu Qiang, Wang Geng, Yi Zhao, Ruibo Fu, Tao Wang, Cheng Gong, Tianrui Wang, Qiuyu Liu, Jiangyan Yi, Zhengqi Wen, et al. 2025. Vq-ctap: Cross-modal fine-grained sequence representation learning for speech processing. *IEEE Transactions on Audio, Speech and Language Processing*.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.

Yao Shi, Hui Bu, Xin Xu, Shaoji Zhang, and Ming Li. 2020. Aishell-3: A multi-speaker mandarin tts corpus and the baselines. *arXiv preprint arXiv:2010.11567*.

Hubert Siuzdak, Florian Grötschla, and Luca A Lanzendörfer. 2024. Snac: Multi-scale neural audio codec. *arXiv preprint arXiv:2410.14411*.

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. 2024. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063.

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

Aaron Van Den Oord, Oriol Vinyals, et al. 2017. Neural discrete representation learning. *Advances in neural information processing systems*, 30.

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

Ho-Hsiang Wu, Prem Seetharaman, Kundan Kumar, and Juan Pablo Bello. 2022. Wav2clip: Learning robust audio representations from clip. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4563–4567. 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*.

Lu Yuan, Dongdong Chen, Yi-Ling Chen, Noel Codella, Xiyang Dai, Jianfeng Gao, Houdong Hu, Xuedong Huang, Boxin Li, Chunyuan Li, et al. 2021. Florence: A new foundation model for computer vision. *arXiv preprint arXiv:2111.11432*.

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*.

Aohan Zeng, Zhengxiao Du, Mingdao Liu, Kedong Wang, Shengmin Jiang, Lei Zhao, Yuxiao Dong, and Jie Tang. 2024. Glm-4-voice: Towards intelligent and human-like end-to-end spoken chatbot. *arXiv preprint arXiv:2412.02612*.

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

Ziqiang Zhang, Long Zhou, Chengyi Wang, Sanyuan Chen, Yu Wu, Shujie Liu, Zhuo Chen, Yanqing Liu, Huaming Wang, Jinyu Li, et al. 2023b. Speak foreign languages with your own voice: Cross-lingual neural codec language modeling. *arXiv preprint arXiv:2303.03926*.

A Related work

A.1 Neural Codec

Discretized vector quantization(VQ-VAE)(Van Den Oord et al., 2017) has become a fundamental component for compression in codecs(Gârbacea et al., 2019). The introduction of residual vector quantization (RVQ)(Zeghidour et al., 2021) enables high-quality audio compression with scalable bitrates. Current codecs are categorized into acoustic coding codecs (such as Encodec(Défossez et al., 2022), SoundStream(Zeghidour et al., 2021), and SNAC(Siuzdak et al., 2024)) and semantic disentanglement codecs (such as FACodec(Ju et al., 2024), SpeechTokenizer(Zhang et al., 2023a), MimiCodec(Défossez et al., 2024), VQ-CTAP(Qiang et al., 2025) and SemantiCodec(Liu et al., 2024)). Many single-codebook codec models are also proposed, such as WavTokenizer(Ji et al., 2024), BigCodec(Xin et al., 2024) and TAAE(Parker et al., 2024). Methods such as FACodec, SpeechTokenizer, and MimiCodec supervise the semantic encoding only in the first-dimensional VQ space, often incorporating ASR tasks(Ganin and Lempitsky, 2015) or distillation tasks from Hubert(Hsu et al., 2021)/ WavLM(Chen et al., 2022). However, the representations from Hubert/ WavLM are not purely semantic and contain significant residual paralinguistic information. These models combine semantic encodings with other encodings during the decoding stage to reconstruct speech. This approach limits the disentanglement and reconstruction capability of the semantic encoding itself. In TTS tasks, there is no requirement for streaming in speech encoding. However, for voice-based dialogue tasks(Défossez et al., 2024), streaming speech encoding is essential. Among the methods mentioned, SoundStream, Encodec, and MimiCodec support streaming encoding and decoding through causal structures.

A.2 Contrastive Learning

Contrastive learning works by differentiating a target sample (positive) from distractor samples (negatives) based on an anchor representation. The objective is to maximize the similarity between the anchor and positive samples while minimizing the similarity between the anchor and negative samples. This approach has been widely applied in the field of computer vision, with notable examples such as Open AI's CLIP(Radford et al., 2021), Florence(Yuan et al., 2021), and ALIGN(Jia et al., 2021). In the audio field, CLIP-based models have also been developed, including Wav2CLIP(Wu et al., 2022), AudioCLIP(Guzhov et al., 2022), and CLAP(Elizalde et al., 2023). These models focus on extracting global descriptive information from audio, with the primary goal of improving the performance of downstream audio classification tasks. The downstream tasks applied in the experimental parts of these studies are primarily speech classification tasks, such as sound event classification, instrument classification, acoustic scene classification, emotion recognition, keyword spotting, vocal sound classification, and speaker counting. Since the global features extracted by these methods lose temporal information, they cannot be converted back into frame-level acoustic features. VQ-CTAP(Qiang et al., 2025) is a frame-level speech representation model based on contrastive learning. However, it does not support streaming encoding and decoding.

B Subjective Evaluation Details

To rigorously assess the perceptual quality of the synthesized speech, we conducted a subjective listening test following the MUSHRA (Multiple Stimuli with Hidden Reference and Anchor) protocol.

Participants. We recruited 20 volunteer listeners. The participants were provided with high-quality headphones and instructed to perform the test in a quiet environment.

Procedure and Instructions. The test was conducted using a web-based MUSHRA interface. Participants were asked to rate the quality of each sample on a scale from 0 to 100, considering factors such as:

- **Signal Quality**: The presence of artifacts, noise, or distortion.
- **Intelligibility**: The clarity and pronunciation of the speech.

Participants were explicitly instructed to rate the hidden reference as 100. We filtered out responses from participants who failed to rate the hidden reference above 95 to ensure data quality. The final MUSHRA scores reported in Table 1 are the average scores across all valid participants and test samples.

C Algorithm and Visualizations

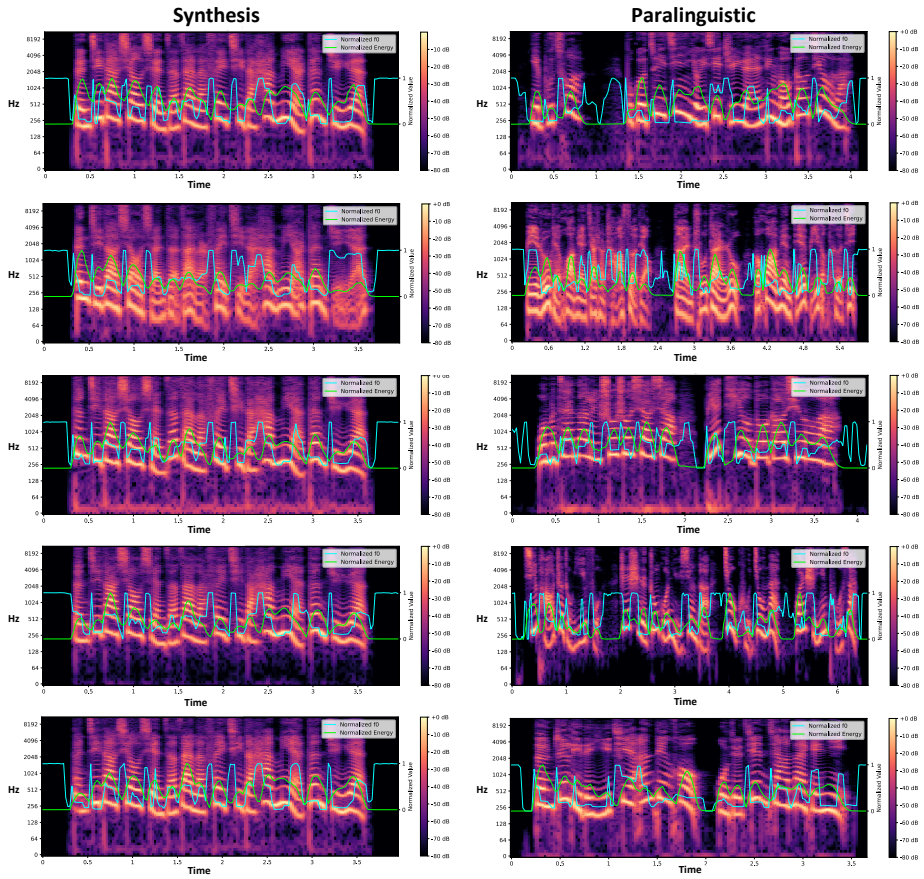
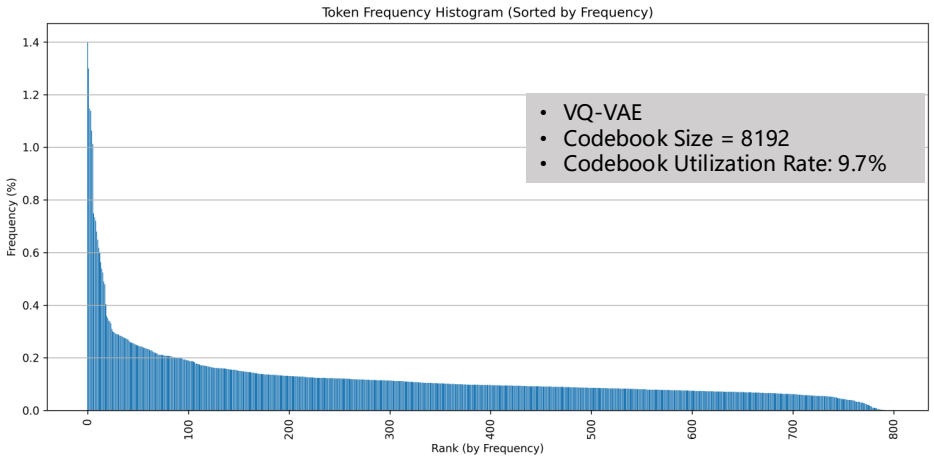
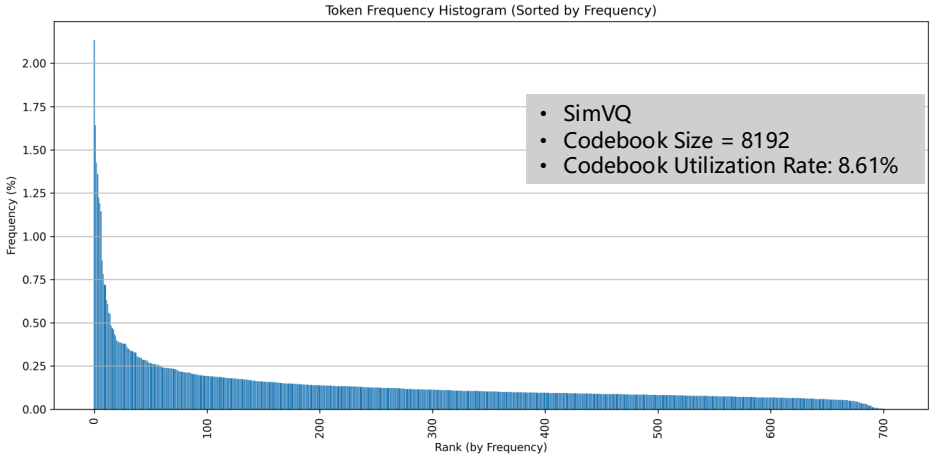


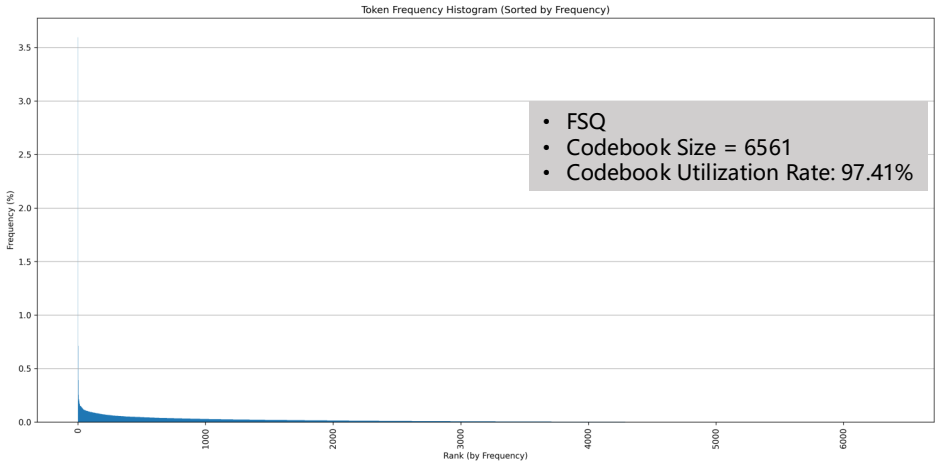
Figure 4: The Spectrograms, F0, and Energy of synthesized speech (the same semantic coding combining different paralinguistics). The bottom row is the ground-truth. The synthesized speech and the paralinguistic speech exhibit consistency in the numerical range and variation trends of spectrogram, F0, and energy.

Algorithm 1 Acoustic-Constrained Multi-Stage Optimization

Initialize $step$, $stage1_{end}$, kl_{start}^{para}, kl_{end}^{para}, kl_{upper}^{para}, $kl_{start}^{semantic}$, $kl_{end}^{semantic}$, $kl_{upper}^{semantic}$, α, β
for each training $step$ **do**
  if $step \leq stage1_{end}$ **then**
    Train Speech Encoder, Acoustic Projection & Speech Decoder with $\mathcal{L}_{mse}^{mel}$
  else
    Freeze modules from Stage 1
    $\mathcal{L}_{total} \leftarrow \alpha \mathcal{L}_{mse}^{acoustic} + \beta \mathcal{L}_{contrastive}$
    if $step > kl_{start}^{para}$ **then**
      $\gamma \leftarrow kl_{upper}^{para} \cdot \min(1, \frac{step - kl_{start}^{para}}{kl_{end}^{para} - kl_{start}^{para}})$
      $\mathcal{L}_{total} \leftarrow \mathcal{L}_{total} + \gamma \mathcal{L}_{kl}^{para}$
    end if
    if $step > kl_{start}^{semantic}$ **then**
      $\delta \leftarrow kl_{upper}^{semantic} \cdot \min(1, \frac{step - kl_{start}^{semantic}}{kl_{end}^{semantic} - kl_{start}^{semantic}})$
      $\mathcal{L}_{total} \leftarrow \mathcal{L}_{total} + \delta \mathcal{L}_{kl}^{semantic}$
    end if
    Train Phoneme Encoder, Paralinguistic Encoder, Semantic Projector & Connector with $\mathcal{L}_{total}$
  end if
end for







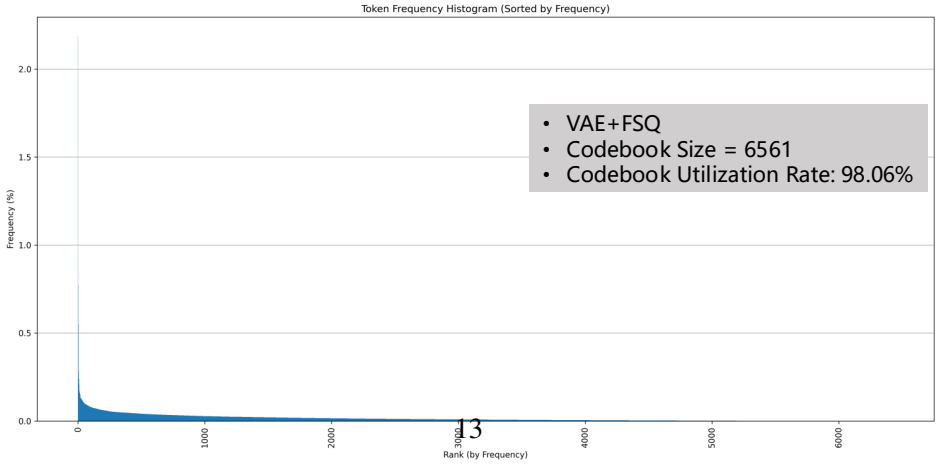


Figure 5: Codebook Utilization.