Spark-TTS: An Efficient LLM-Based Text-to-Speech Model with Single-Stream Decoupled Speech Tokens

Anonymous ACL submission

Abstract

Recent advances in large language models (LLMs) have enabled remarkable progress in zero-shot text-to-speech (TTS) synthesis, yet existing foundation models face significant limitations. While these models excel at reproducing voices from reference audio, they lack fine-grained control over voice attributes and, in single-stream approaches, suffer from the entanglement of semantic and acoustic information within tokens. This entanglement makes independent manipulation of speech characteristics challenging and hinders the creation of entirely new voices. To address these limitations, we introduce Spark-TTS, a novel system built upon our proposed BiCodec, a single-stream speech codec that strategically decomposes speech into two complementary token types: low-bitrate semantic tokens for linguistic content and fixed-length global tokens for speaker-specific attributes. This disentangled representation, combined with the Qwen2.5 LLM and a chain-of-thought (CoT) generation approach, enables both coarse-grained attribute control (e.g., gender, speaking style) and fine-grained parameter adjustment (e.g., precise pitch values, speaking rate). To advance research in controllable TTS, we introduce VoxBox, a meticulously curated 100,000-hour dataset with comprehensive attribute annotations. Extensive experiments demonstrate that Spark-TTS not only achieves state-of-the-art performance in zero-shot voice cloning but also excels at generating novel, highly customizable voices that transcend the limitations of reference-based synthesis[1]. Audio samples are available at `https://spark-tts.github.io/`.

1 Introduction

Recent advances in speech tokenization have revolutionized text-to-speech (TTS) synthesis by bridging the fundamental gap between continuous speech signals and discrete token-based large language models (LLMs) (Anastassiou et al., 2024; Zhu et al., 2024; Wang et al., 2024c). Through sophisticated quantization techniques, particularly Vector Quantization (VQ) (Van Den Oord et al., 2017) and Finite Scalar Quantization (FSQ) (Mentzer et al., 2023), codec-based LLMs have emerged as the predominant paradigm for zero-shot TTS. The integration of extensive training data with large-scale model architectures has enabled these systems to achieve unprecedented levels of naturalness, often rendering synthetic speech indistinguishable from human speech (Anastassiou et al., 2024; Du et al., 2024b; Chen et al., 2024b; Ye et al., 2024a).

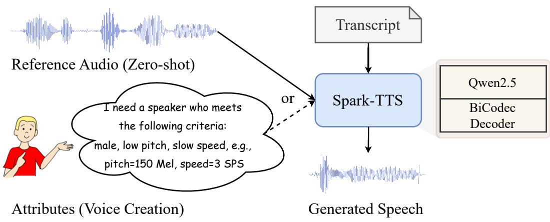


Figure 1: Spark-TTS can generate in a zero-shot manner through reference audio, as well as create new speakers by leveraging coarse- or fine-grained attribute control.

Despite the remarkable progress in LLM-based zero-shot TTS, several fundamental challenges persist. Current codec-based TTS architectures exhibit significant complexity, requiring either dual generative models (Wang et al., 2023a; Anastassiou et al., 2024) or intricate parallel multi-stream code prediction mechanisms (Kreuk et al., 2023; Le Lan et al., 2024) that deviate substantially from conventional text LLM frameworks. This divergence stems from inherent limitations in existing audio codecs - while semantic tokens provide compactness, they necessitate additional models for acoustic feature prediction (Du et al., 2024a; Huang et al., 2023) and lack integrated timbre control capabilities. Acoustic tokens, meanwhile, rely on complex codebook architectures like group-VQ (Défossez et al., 2022; Van Den Oord et al., 2017). The field also struggles

[1] Source code and checkpoint will be released.

with the creation of novel voices, as current systems are predominantly limited to reference-based generation (Zhang et al., 2023b; Chen et al., 2024a), lacking the capability to synthesize voices with precisely specified characteristics. This limitation is further compounded by insufficient granularity in attribute control, especially for fine-grained characteristics such as pitch modulation, despite recent advances in instruction-based generation (Du et al., 2024b). Furthermore, the prevalent use of proprietary datasets in current research creates significant challenges for standardized evaluation and meaningful comparison of methods (Anastassiou et al., 2024; Ye et al., 2024a). These limitations collectively underscore the need for a unified approach that can simplify architecture, enable flexible voice creation with comprehensive attribute control, and establish reproducible benchmarks through open data resources.

To address these fundamental limitations, we introduce Spark-TTS, a unified system that achieves zero-shot TTS with comprehensive attribute control through a single codec LLM, maintaining architectural alignment with conventional text LLMs. In addition, we present VoxBox, a meticulously curated and annotated open-source speech dataset that establishes a foundation for reproducible research in speech synthesis. Specifically, we introduce BiCodec, a novel tokenization framework that preserves the efficiency of semantic tokens while enabling fine-grained control over timbre-related attributes. BiCodec achieves this through combining low-bitrate semantic tokens with fixed-length global tokens, effectively capturing both linguistic content and time-invariant acoustic characteristics. Building upon BiCodec, we leverage Qwen2.5 (Yang et al., 2024) through targeted fine-tuning, seamlessly integrating TTS capabilities within the text LLM paradigm. To enable comprehensive voice control, we implement a hierarchical attribute system combining coarse-grained labels (gender, pitch, speaking speed) with fine-grained numerical values, orchestrated through a chain-of-thought (CoT) prediction framework.

Our primary contributions encompass:

- **New Tokenization**: We present BiCodec, a unified speech tokenization that generates a hybrid token stream combining semantic and global tokens. This approach maintains linguistic fidelity while enabling sophisticated attribute control through LM-based mechanisms.
- **Coarse- and Fine-Grained Voice Control**: Spark-TTS implements a comprehensive attribute control system that seamlessly integrates both categorical and continuous parameters within a text LLM-compatible architecture. As demonstrated in Fig. 1, this innovation transcends traditional reference-based approaches to zero-shot TTS.
- **Benchmark Dataset**: We introduce VoxBox, a rigorously curated 100,000-hour speech corpus, developed through systematic data collection, cleaning, and attribute annotation. This resource establishes a standardized benchmark for TTS research and evaluation.

2 Related Work

2.1 Single-Stream Speech Tokenizer

Early single-stream speech tokenizers primarily focused on extracting semantic tokens (Huang et al., 2023; Du et al., 2024a; Tao et al., 2024). While pure semantic tokens enable low-bitrate encoding, they necessitate an additional acoustic feature prediction module in semantic token-based speech synthesis (Du et al., 2024a,b).

Recently, single-stream-based acoustic tokenization has gained considerable attention (Xin et al., 2024; Wu et al., 2024). WavTokenizer (Ji et al., 2024a) employs a convolution-based decoder to improve reconstruction quality, while X-codec2 (Ye et al., 2025) enlarges the code space with FSQ. Instead of following a pure encoder-VQ-decoder paradigm, decoupling speech content has proven effective in reducing bitrate using a single codebook (Li et al., 2024a; Zheng et al., 2024).

Among these methods, TiCodec (Ren et al., 2024) is the most similar to our approach in handling global information. However, unlike TiCodec, the proposed BiCodec employs semantic tokens as its time-variant tokens. Instead of using group GVQ (Ren et al., 2024), we propose a novel global embedding quantization method based on FSQ with learnable queries and a cross-attention mechanism. This approach enables the generation of a relatively longer token sequence, offering a more expressive and flexible representation.

2.2 LLM-based Zero-Shot TTS

Prevalent codec LLMs zero-shot TTS predominantly fall into two categories. The first type in-

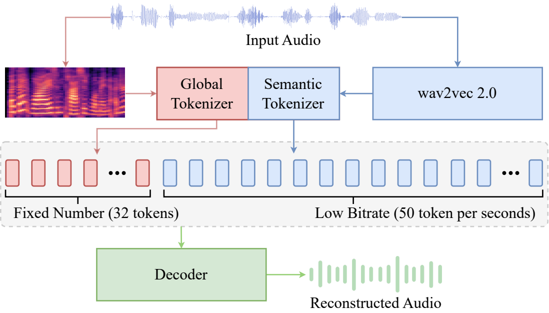


Figure 2: Illustration of the BiCodec. The Global Tokenizer processes the Mel spectrogram to produce global tokens with fixed length, while the Semantic Tokenizer adopts features from wav2vec 2.0 to produce 50 TPS semantic tokens. The decoder reconstructs the waveform from the generated tokens. The detailed structure of BiCodec is provided in Appendix A.

volves predicting single-stream codes using LLMs, followed by the generation of codes enriched with detailed acoustic or continuous semantic features through another LLM (Zhang et al., 2023b; Chen et al., 2024a; Wang et al., 2024a) or generative diffusion models (Anastassiou et al., 2024; Casanova et al., 2024). The second type involves predicting multi-stream codes using carefully designed parallel strategies (Le Lan et al., 2024; Copet et al., 2024) or masked generative patterns (Garcia et al., 2023; Ziv et al., 2024; Li et al., 2024b).

By leveraging the single-stream tokens produced by the proposed BiCodec, Spark-TTS simplifies the modeling of speech tokens within an LLM framework that is fully unified with text LLMs. The most comparable work is the concurrent TTS model Llasa (Ye et al., 2025), which employs an FSQ-based tokenizer to encode speech into single-stream codes with a codebook size of 65,536, followed by LLaMA (Touvron et al., 2023) for speech token prediction. In contrast, Spark-TTS extends beyond zero-shot TTS by integrating speaker attribute labels, enabling controllable voice creation. Additionally, Spark-TTS achieves higher zero-shot TTS performance while using fewer model parameters, enhancing both efficiency and flexibility.

3 BiCodec

To achieve both the compact nature and semantic relevance of semantic tokens, while also enabling acoustic attribute control within an LM, we propose BiCodec, which discretizes input audio into: (i) Semantic tokens at 50 tokens per second (TPS), capturing linguistic content, and (ii) Fixed-length global tokens, encoding speaker attributes and other global speech characteristics.

3.1 Overview

As shown in Fig. 2, BiCodec includes a Global Tokenizer and a Semantic Tokenizer. The former extracts global tokens from the Mel spectrogram of input audio. The latter uses features from wav2vec 2.0 (Baevski et al., 2020) as input to extract semantic tokens.

The BiCodec architecture follows a standard VQ-VAE encoder-decoder framework, augmented with a global tokenizer. The decoder reconstructs discrete tokens back into audio. For an input audio signal $\boldsymbol{x} \in [-1, 1]^T$, with sample number of T, BiCodec functions as follows:

$$\begin{aligned} \boldsymbol{z} &= E_s(F(\boldsymbol{x})), \boldsymbol{g} = E_g(\mathrm{Mel}(\boldsymbol{x})), \\ \boldsymbol{g}_f &= \mathrm{CrossAttention}(\boldsymbol{g}, \boldsymbol{h}), \\ \boldsymbol{z}_q &= Q_s(\boldsymbol{z}), \boldsymbol{g}_q = Q_g(\boldsymbol{g}_f), \\ \hat{\boldsymbol{x}} &= G(\boldsymbol{z}_q, A_g(\boldsymbol{g}_q)), \end{aligned} \tag{1}$$

where $E_s(\cdot)$ is the encoder of the semantic tokenizer, $F(\cdot)$ is the pre-trained wav2vec 2.0[2], $E_g(\cdot)$ is the encoder of the global tokenizer, $\mathrm{Mel}(\cdot)$ is to extract Mel spectrogram from $\boldsymbol{x}$, $\boldsymbol{h}$ is a sequence of learnable queries matching the length of the final global token sequence, $Q_s(\cdot)$ is a quantization layer with VQ, $Q_g(\cdot)$ is a quantization layer with FSQ, $A_g(\cdot)$ is an aggregation module with a pooling layer, and $G(\cdot)$ is the decoder that reconstructs the time-domain signal $\hat{\boldsymbol{x}}$.

3.2 Model Structure

Encoder and Decoder The encoder of the semantic tokenizer E_s and the decoder G are fully convolutional neural networks built with ConvNeXt (Liu et al., 2022) blocks. To effectively capture semantic information, based on the relationship between different layer features of wav2vec 2.0 (XLSR-53) and semantics (Pasad et al., 2023), we select features from the 11th, 14th, and 16th layers, averaging them to obtain the semantic feature, which serves as the input for the semantic tokenizer. The features from the first two layers show a strong correlation with words, while the features from the 16th layer exhibit the strongest correlation with phonemes.

The global tokenizer's encoder, E_g, uses the ECAPA-TDNN architecture (Desplanques et al., 2020) following the implementation by Wespeaker (Wang et al., 2023b) up to the final pooling

[2] https://huggingface.co/facebook/wav2vec2-large-xlsr-53

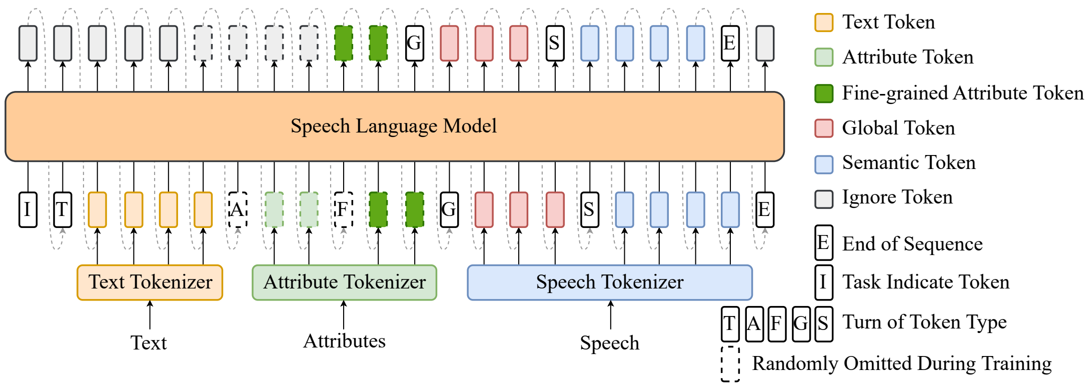


Figure 3: Speech language model of Spark-TTS. During inference, if the input contains attribute tokens representing gender, pitch level, and speed level, the model can predict the corresponding fine-grained attribute tokens, global tokens, and semantic tokens without requiring reference audio in a CoT manner. Otherwise, global tokens can be derived from the reference audio for zero-shot TTS.

layer. After encoding, the global tokenizer extracts a fixed-length sequence representation $\boldsymbol{g}_f$ using a cross-attention mechanism with a set of learnable queries.

Quantization The semantic tokenizer employs single-codebook vector quantization for quantization. Inspired by DAC (Kumar et al., 2024), we use factorized codes to project the encoder's output into a low-dimensional latent variable space prior to quantization.

Considering that the global tokenizer requires a set of discrete tokens to represent time-independent global information, FSQ is employed rather than VQ to mitigate the potential risk of training collapse associated with VQ. Details about the model structure can be seen in Appendix A.

3.3 Training objective

Loss Functions BiCodec is trained end-to-end employing a Generative Adversarial Network (GAN) methodology (Goodfellow et al., 2020) to minimize reconstruction loss, together with L1 feature matching loss (via discriminators) (Kumar et al., 2019, 2024) while simultaneously optimizing the VQ codebook.

Following (Kumar et al., 2024), we compute the frequency domain reconstruction loss using L1 loss on multi-scale mel-spectrograms. Multi-period discriminator (Kong et al., 2020; Engel et al., 2020; Gritsenko et al., 2020) and multi-band multi-scale STFT discriminator (Kumar et al., 2024) are used for waveform discrimination and frequency domain discrimination, respectively.

VQ codebook learning incorporates both a codebook loss and a commitment loss. Following the approach in (Xin et al., 2024), the codebook loss is calculated as the L1 loss between the encoder output and the quantized results, employing stop-gradients. Additionally, the straight-through estimator (Bengio et al., 2013) is used to enable the backpropagation of gradients.

To ensure training stability, in the initial stages, the global embedding derived from the averaged $\boldsymbol{g}_q$ is not integrated into the decoder. Instead, this embedding is obtained directly from the pooling of $\boldsymbol{g}_f$. Meanwhile, the FSQ codebook is updated using an L1 loss between embedding obtained from $\boldsymbol{g}_f$ and that from $\text{pool}(\boldsymbol{g}_q)$. As training progresses and stabilizes, this teacher-student form will be omitted after a specific training step.

To further ensure semantic relevance, following X-Codec (Ye et al., 2024b), a wav2vec 2.0 reconstruction loss is applied after quantization, with ConvNeXt-based blocks serving as the predictor.

4 Language Modeling of Spark-TTS

4.1 Overview

As illustrated in Fig. 3, the Spark-TTS speech language model adopts a decoder-only transformer architecture, unified with a typical textual language model. We employ the pre-trained textual LLM Qwen2.5-0.5B[3] (Yang et al., 2024) as the backbone of the speech language model. Unlike CosyVoice2 (Du et al., 2024a), Spark-TTS does not require flow matching to generate acoustic features. Instead, BiCodec's decoder directly processes the LM's output to produce the final audio, significantly simplifying the textual LLM-based speech generation pipeline.

[3] https://huggingface.co/Qwen/Qwen2.5-0.5B-Instruct

In addition to zero-shot TTS, Spark-TTS supports voice creation using various attribute labels. During inference, if attribute labels for gender, pitch level, and speed level are provided, the language model can predict fine-grained pitch values, speed values, global tokens, and semantic tokens through chain-of-thought processing. If no attribute labels are provided, global tokens are extracted from the reference audio, enabling zero-shot TTS.

4.2 Tokenizer

Text Tokenizer Similar to textual LLMs, Spark-TTS employs a byte pair encoding (BPE)-based tokenizer to process raw text. Here, we adopt the Qwen2.5 tokenizer (Yang et al., 2024), which supports multiple languages.

Attribute Tokenizer To enable voice creation based on speech attributes, Spark-TTS encodes attribute information at two levels: (i) *Coarse-Grained*: Attribute labels representing high-level speech characteristics, including gender, pitch (categorized into five discrete levels), and speed (categorized into five discrete levels); (ii) *Fine-Grained*: Attribute values enabling precise control over pitch and speed, which are quantized by rounding to the nearest integer during tokenization.

Speech Tokenizer The speech tokenizer consists of a global tokenizer and a semantic tokenizer. Using both global and semantic tokens, the BiCodec decoder reconstructs the waveform signal.

4.3 Training Objective

The decoder-only language model is trained by minimizing the negative log-likelihood of token predictions. Let $\mathcal{T}$ represent the tokenized textual prompt and $\mathcal{G}$ denote the global speech token prompt; the optimization for zero-shot TTS is defined as follows:

$$\mathcal{L}_{zst} = -\sum_{t=1}^{T_o} \log P(o_t|\mathcal{T}, \mathcal{G}, \boldsymbol{o}_{<t}; \theta_{LM}), \quad (2)$$

where $\boldsymbol{o} \in \mathbb{N}_o^T$ represents the semantic tokens to be predicted in the zero-shot TTS scenario, and θ_{LM} denotes the parameters of the language model.

For the case of voice creation, the optimization is defined as follows:

$$\mathcal{L}_{control} = -\sum_{t=1}^{T_c} \log P(c_t|\mathcal{T}, \mathcal{A}, \boldsymbol{c}_{<t}; \theta_{LM}), \quad (3)$$

where $\mathcal{A}$ represents the attribute label prompt, and the output $\boldsymbol{c}$ encompasses $\mathcal{F}$, $\mathcal{G}$, and $\mathcal{S}$. Here, $\mathcal{F}$ denotes the fine-grained attribute value prompt, and $\mathcal{S}$ is speech semantic tokens.

In practice, $\mathcal{L}_{zst}$ and $\mathcal{L}_{control}$ are mixed during training. Specifically, each audio example is structured into two training samples according to $\mathcal{L}_{zst}$ and $\mathcal{L}_{control}$ respectively.

5 VoxBox

5.1 Overview

To facilitate voice creation and establish a fair comparison benchmark for future research, we introduce VoxBox, a well-annotated dataset for both English and Chinese. All data sources in VoxBox originate from open-source datasets, ensuring broad accessibility. To enhance data diversity, we collect not only common TTS datasets, but also datasets used for speech emotion recognition. Each audio file in VoxBox is annotated with gender, pitch, and speed. Additionally, we also perform data cleaning on datasets with lower text quality. After data cleaning, VoxBox comprises 4.7 million audio files, sourced from 29 open datasets, totaling 102.5k hours of speech data. Details about VoxBox and the source datasets can be found in Appendix E.

5.2 Clean and Annotation

Gender Annotation Given the strong performance of pre-trained WavLM in speaker-related tasks (Li et al., 2024c), we fine-tune the WavLM-large model for gender classification using datasets that contain explicit gender labels (detailed in Appendix E.2). Our fine-tuned model achieves 99.4% accuracy on the AISHELL-3 test set. We then use this gender classification model to annotate datasets previously lacking gender labels.

Pitch Annotation We extract the average pitch value from each audio clip using PyWorld[4], rounding it to the nearest integer to obtain fine-grained pitch value tokens. For the definition of pitch levels, we first convert the average pitch of each audio clip to the Mel scale. We then conduct a statistical analysis of all Mel scale pitch for all males and females separately. Based on the 5th, 20th, 70th, and 90th percentiles, we establish boundaries for five pitch levels: very low, low, moderate, high, and very high (detailed in Appendix E.1).

Speed Annotation Compared to character-based (Vyas et al., 2023), word-based (Ji et al., 2024b), or phoneme-based (Lyth and King, 2024)

[4] https://pypi.org/project/pyworld/

speaking rate calculations, syllable-based measurements provide a more direct correlation with speaking rate. Here, we initially apply Voice Activity Detection (VAD) to eliminate silent segments at both ends. Subsequently, we calculate the syllables per second (SPS), which is then rounded to the nearest integer to serve as the fine-grained speed value token. Using the 5th, 20th, 80th, and 95th percentiles, we establish boundaries for five distinct speed levels: very slow, slow, moderate, fast, and very fast (detailed in Appendix E.1).

Data Cleaning For datasets exhibiting lower text quality, we conduct an additional cleaning process. Specifically, for Emilia (He et al., 2024), the original transcripts were obtained using the Whisper-based (ASR) system (Radford et al., 2023), employing the whisper-medium model, which occasionally resulted in inaccuracies. To address this, we employ another ASR model, FunASR (Gao et al., 2023)[5], to re-recognize the audio. We then use the original scripts as ground truth to calculate the Word Error Rate (WER) and excluded samples with a WER exceeding 0.05. For the MLS-English, LibriSpeech, LibriTTS-R, and datasets originally designed for emotion recognition, we employ the whisper-large-v3[6] model for speech recognition, comparing the recognition results with the original scripts. Samples exhibiting insertions or deletions are excluded from the dataset.

6 Experiments

6.1 Implementation Details

BiCodec is trained on the full training set of the LibriSpeech dataset, comprising 960 hours of English speech data. Additionally, we include 1,000 hours of speech data from both Emilia-CN and Emilia-EN, bringing the total training data to approximately 3,000 hours. All audio samples are resampled to 16 kHz. The global token length is set as 32. For optimization, we use the AdamW optimizer with moving average coefficients coefficients $\beta_1 = 0.8$ and $\beta_2 = 0.9$. The model converges within approximately 800k training steps using a batch size with 614.4 seconds of speech.

The Spark-TTS language model is trained using the entire VoxBox training set. If a dataset lacks predefined train/test splits, we use the entire processed dataset for training. The training employs the AdamW optimizer with $\beta_1 = 0.9$ and $\beta_2 = 0.96$. The model undergoes training over 3 epochs, using a batch size of 768 samples.

[5]ZH: https://huggingface.co/funasr/paraformer-zh
EN: https://huggingface.co/FunAudioLLM/SenseVoiceSmall

[6]https://huggingface.co/openai/whisper-large-v3

6.2 Reconstruction Performance of BiCodec

Comparsion with Other Methods The reconstruction performance of BiCodec compared to other methods is presented in Table 1. As can be seen, within the low-bitrate range (<1 kbps), BiCodec surpasses all methods on most metrics, except for UTMOS, where it ranks second to StableCodec, and SIM, where it ranks second to X-Codec2, thereby achieving a new state-of-the-art (SOTA) performance.

Notably, BiCodec's semantic tokens are extracted from wav2vec 2.0 rather than raw audio, resulting in stronger semantic alignment compared to codecs that directly process waveform-based representations. Further experimental results and analyses are provided in Appendix A.3.

Effectiveness of Global Tokenizer We first evaluate the optimal length for the global token sequence. As shown in Table 2, we compare the impact of different sequence lengths on reconstruction quality. The results without FSQ quantization serve as a benchmark reference. Notably, increasing the global token sequence length consistently improves reconstruction quality, with performance approaching the benchmark at the length of 32.

Furthermore, Table 2 compares our proposed quantization method—which incorporates learnable queries and FSQ—against the GVQ-based method introduced by Ren et al. (Ren et al., 2024) for time-invariant codes. Our approach demonstrates a substantial performance improvement over the GVQ-based method, highlighting the effectiveness of FSQ with learnable queries in enhancing global token representation.

6.3 Control Capabilities of Spark-TTS

Spark-TTS enables controllable generation by inputting attribute labels or fine-grained attribute values. In label-based control, the model automatically generates the corresponding attribute values (e.g., pitch and speed). However, when these values are manually specified, the system switches to fine-grained control.

Gender To assess Spark-TTS's capability in gender control, we compare it with textual prompt-based controllable TTS models, including VoxInstruct(Zhou et al., 2024b) and Parler-TTS(Lyth and King, 2024). For evaluation, we reorganize the

Table 1: Comparisons of various codec models for speech reconstruction on the LibriSpeech test-clean dataset. Detailed information about these models can be found in Appendix A.2.

Model	Codebook Size	Nq	Token Rate (TPS)	Bandwidth (bps)	STOI↑	PESQ NB↑	PESQ WB↑	UTMOS↑	SIM↑
Encodec	1024	8	600	6000	0.94	3.17	2.75	3.07	0.89
DAC	1024	12	600	6000	**0.95**	**4.15**	**4.01**	4.00	**0.98**
Encodec	1024	2	150	1500	0.84	1.94	1.56	1.58	0.6
Mimi	2048	8	100	1100	0.91	2.8	2.25	3.56	0.73
BigCodec	8192	1	80	1040	0.94	3.27	2.68	4.11	0.84
DAC	1024	2	100	1000	0.73	1.4	1.14	1.29	0.32
SpeechTokenizer	1024	2	100	1000	0.77	1.59	1.25	2.28	0.36
X-codec	1024	2	100	1000	0.86	2.88	2.33	4.21	0.72
WavTokenizer	4096	1	75	900	0.89	2.64	2.14	3.94	0.67
X-codec2	65536	1	50	800	0.92	3.04	2.43	4.13	**0.82**
StableCodec	15625	2	50	697	0.91	2.91	2.24	**4.23**	0.62
Single-Codec	8192	1	23.4	304	0.86	2.42	1.88	3.72	0.60
BiCodec	8192	1	50	650	**0.92**	**3.13**	**2.51**	4.18	0.80

Table 2: Performance of BiCodec with varying global token lengths for reconstruction on the LibriSpeech test-clean dataset, where "w/o" indicates the omission of FSQ-based quantization, and gvq-32 means the global tokenizer is implemented with group VQ. For performance results on the LibriTTS test-clean dataset, refer to Appendix A.3.

Global Token	STOI↑	PESQ NB↑	PESQ WB↑	UTMOS↑	SIM↑
w/o FSQ	0.915	**3.14**	**2.52**	4.15	**0.83**
gvq-32	0.912	2.91	2.30	4.06	0.74
8	0.916	3.04	2.41	4.16	0.74
16	0.919	3.08	2.45	4.15	0.77
32	**0.922**	3.13	2.51	**4.18**	0.80

Table 3: Gender control performance of various models.

Method	VoxInstruct	Parler-tts	Spark-TTS
Acc (%)↑	82.99	98.12	**99.77**

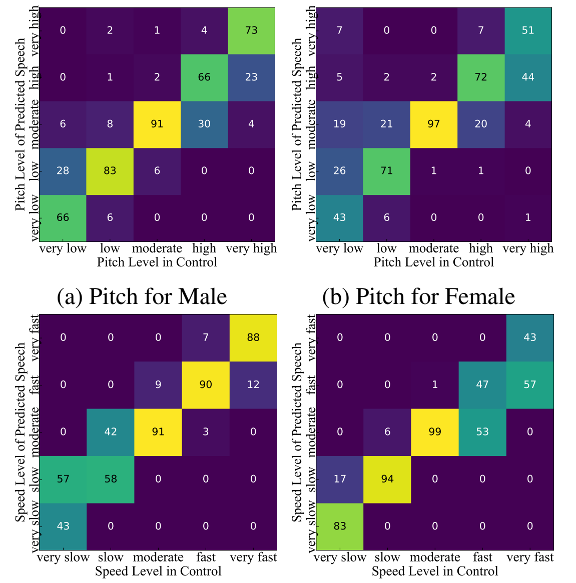
test prompts of real speech from PromptTTS (Guo et al., 2023) based on the prompt structures used in VoxInstruct and Parler-TTS. The gender accuracy (Acc) of the generated speech is measured using our gender predictor, which is specifically trained for gender annotation. The results, presented in Table 3, show that Spark-TTS significantly outperforms other controllable TTS systems in gender control, demonstrating its strong capability in attribute-based voice generation.

Pitch and Speed Spark-TTS enables controllable generation by inputting attribute labels or fine-grained attribute values. In label-based control, the model automatically generates the corresponding attribute values (e.g., pitch and speed). However, when these values are manually specified, the system switches to fine-grained control. Fig. 4 illustrates the control confusion matrices for pitch and speaking rate based on coarse-grained labels, while Fig. 5 presents the fine-grained control performance for pitch and speed. As shown, Spark-TTS accurately generates speech that aligns with the specified attribute labels, demonstrating precise control over both coarse-grained and fine-grained attributes.

6.4 Zero-shot TTS Performance

To evaluate Spark-TTS's zero-shot TTS capability, we assess its performance on Seed-TTS-eval and compare it with existing zero-shot TTS models. The results are presented in Table 4, where speech intelligibility is evaluated using the Character Error Rate (CER) for Chinese and the WER for English, following the Seed-TTS-eval[7]. As can been seen, Spark-TTS demonstrates significant superiority in intelligibility for zero-shot TTS scenarios. On test-zh, Spark-TTS achieves a CER second only to the closed-source model Seed-TTS, while it ranks second only to F5-TTS (Chen et al., 2024b) for English WER. This high intelligibility is partly attributed to the semantic feature-based BiCodec and further validates the high quality of our VoxBox dataset in terms of transcripts. In terms of speaker similarity, while Spark-TTS is relatively weaker than multi-stage or NAR-based methods, it significantly outperforms the single-stage model Llasa (Ye et al., 2025). Notably, Spark-TTS, with just 0.5B model parameters and 100k hours of training data, surpasses Llasa, which has 8B parameters and is trained on 250k hours of data.

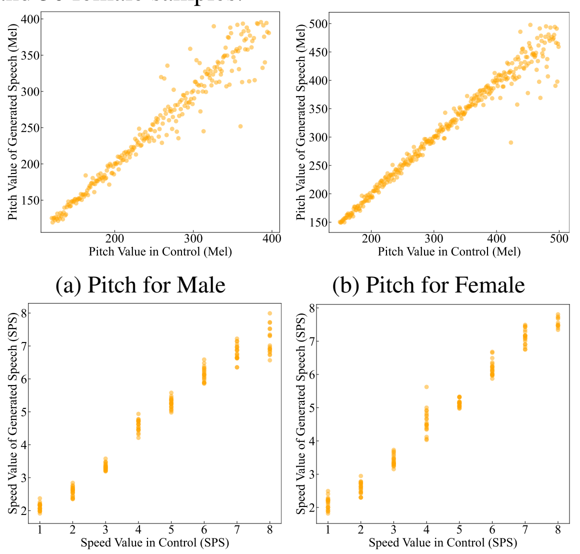
[7] https://github.com/BytedanceSpeech/seed-tts-eval



(a) Pitch for Male (b) Pitch for Female

(c) Speed for English (d) Speed for Chinese

Figure 4: Confusion matrix of coarse-grained pitch and speed control results. In pitch-controllable generation, each label's generated samples consist of 50 Chinese and 50 English samples. In speed-controllable generation, each label's generated samples consist of 50 male and 50 female samples.



(a) Pitch for Male (b) Pitch for Female

(c) Speed for English (d) Speed for Chinese

Figure 5: Fine-grained pitch and speed control results. For pitch-controllable generation, each generated value includes one Chinese sample and one English sample. For speed-controllable generation, each generated value includes 10 male samples and 10 female samples.

Following CosyVoice2 (Du et al., 2024b), we evaluate the quality of the generated speech on the LibriSpeech test-clean set. As shown in Table 5, our method produces audio of significantly higher quality than the original and outperforms CosyVoice2, the SOTA open-source TTS model with multi-stage modeling. This demonstrates the strong performance of Spark-TTS in terms of speech quality.

Table 4: Results of Spark-TTS and recent TTS models on the Seed test sets (test-zh for Chinese and test-en for English). † denotes closed-sourced models.

Model	test-zh		test-en	
	CER↓	SIM↑	WER↓	SIM↑
Multi-Stage or NAR Methods				
Seed-TTS†	**1.12**	**0.796**	2.25	**0.762**
FireRedTTS	1.51	0.635	3.82	0.460
MaskGCT	2.27	0.774	2.62	0.714
E2 TTS (32 NFE)†	1.97	0.730	2.19	0.710
F5-TTS (32 NFE)	1.56	0.741	**1.83**	0.647
CosyVoice	3.63	0.723	4.29	0.609
CosyVoice2	1.45	0.748	2.57	0.652
One-Stage AR Methods				
Llasa-1B-250k	1.89	0.669	3.22	0.572
Llasa-3B-250k	1.60	0.675	3.14	0.579
Llasa-8B-250k	1.59	**0.684**	2.97	0.574
Spark-TTS	**1.20**	0.672	**1.98**	**0.584**

Table 5: Quality comparison of zero-shot TTS audio generation on the LibriSpeech test-clean set. GT represents ground truth.

Method	GT	CosyVoice	CosyVoice2	Spark-TTS
UTMOS↑	4.08	4.09	4.23	**4.35**

7 Conclusion

This paper introduces BiCodec, which retains the advantages of semantic tokens, including high compression efficiency and high intelligibility, while addressing the limitation of traditional semantic tokens, which cannot control timbre-related attributes within an LM, by incorporating global tokens. BiCodec achieves a new SOTA reconstruction quality, operating at 50 TPS with a bit rate of 0.65 kbps, surpassing other codecs within the sub-1 kbps range. Building on BiCodec, we develop Spark-TTS, a text-to-speech model that integrates the textual language model Qwen2.5. Spark-TTS enables voice generation based on specified attributes and supports zero-shot synthesis. To our knowledge, this is the first TTS model to offer fine-grained control over both pitch and speaking rate, while simultaneously supporting zero-shot TTS. Additionally, to facilitate comparative research, we introduce VoxBox, an open-source dataset designed for controllable speech synthesis. VoxBox not only filters out low-quality textual data but also provides comprehensive annotations, including gender, pitch, and speaking rate, significantly enhancing training for controlled generation tasks.

Limitation

Despite its advantages, Spark-TTS also has notable limitations. Similar to Llasa (Ye et al., 2025), which relies on a single codebook and a textual language model, Spark-TTS exhibits relatively lower speaker similarity metrics in zero-shot TTS compared to multi-stage or NAR methods. This may be due to the greater speaker variability introduced by the AR language model during inference. Currently, Spark-TTS does not impose additional disentanglement constraints between global tokens and semantic tokens. In future work, we aim to enhance global token control over timbre by introducing perturbations to formants or pitch in the semantic token input. This approach will promote better disentanglement of timbre information, allowing BiCodec's decoder to exert absolute control over timbre. By doing so, we aim to reduce randomness introduced by the AR model, improving the speaker similarity in zero-shot synthesis.

References

Adaeze Adigwe, Noé Tits, Kevin El Haddad, Sarah Ostadabbas, and Thierry Dutoit. 2018. The emotional voices database: Towards controlling the emotion dimension in voice generation systems. *arXiv preprint arXiv:1806.09514*.

Philip Anastassiou, Jiawei Chen, Jitong Chen, Yuanzhe Chen, Zhuo Chen, Ziyi Chen, Jian Cong, Lelai Deng, Chuang Ding, Lu Gao, et al. 2024. Seed-tts: A family of high-quality versatile speech generation models. *arXiv preprint arXiv:2406.02430*.

Asger Heidemann Andersen, Jan Mark de Haan, Zheng-Hua Tan, and Jesper Jensen. 2017. A non-intrusive short-time objective intelligibility measure. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5085–5089. IEEE.

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

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33:12449–12460.

Evelina Bakhturina, Vitaly Lavrukhin, Boris Ginsburg, and Yang Zhang. 2021. Hi-Fi Multi-Speaker English TTS Dataset. *arXiv preprint arXiv:2104.01497*.

Yoshua Bengio, Nicholas Léonard, and Aaron Courville. 2013. Estimating or propagating gradients through stochastic neurons for conditional computation. *arXiv preprint arXiv:1308.3432*.

Felix Burkhardt, Johannes Wagner, Hagen Wierstorf, Florian Eyben, and Björn Schuller. 2023. Speech-based age and gender prediction with transformers. In *Speech Communication; 15th ITG Conference*, pages 46–50. VDE.

Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Narayanan. 2008. Iemocap: Interactive emotional dyadic motion capture database. *Language resources and evaluation*, 42:335–359.

Houwei Cao, David G Cooper, Michael K Keutmann, Ruben C Gur, Ani Nenkova, and Ragini Verma. 2014. Crema-d: Crowd-sourced emotional multimodal actors dataset. *IEEE transactions on affective computing*, 5(4):377–390.

Edresson Casanova, Kelly Davis, Eren Gölge, Görkem Göknar, Iulian Gulea, Logan Hart, Aya Aljafari, Joshua Meyer, Reuben Morais, Samuel Olayemi, et al. 2024. Xtts: a massively multilingual zero-shot text-to-speech model. *arXiv preprint arXiv:2406.04904*.

Guoguo Chen, Shuzhou Chai, Guanbo Wang, Jiayu Du, Wei-Qiang Zhang, Chao Weng, Dan Su, Daniel Povey, Jan Trmal, Junbo Zhang, et al. 2021. Gigaspeech: An evolving, multi-domain asr corpus with 10,000 hours of transcribed audio. *arXiv preprint arXiv:2106.06909*.

Sanyuan Chen, Shujie Liu, Long Zhou, Yanqing Liu, Xu Tan, Jinyu Li, Sheng Zhao, Yao Qian, and Furu Wei. 2024a. Vall-e 2: Neural codec language models are human parity zero-shot text to speech synthesizers. *arXiv preprint arXiv:2406.05370*.

Yushen Chen, Zhikang Niu, Ziyang Ma, Keqi Deng, Chunhui Wang, Jian Zhao, Kai Yu, and Xie Chen. 2024b. F5-tts: A fairytaler that fakes fluent and faithful speech with flow matching. *arXiv preprint arXiv:2410.06885*.

Zhengyang Chen, Sanyuan Chen, Yu Wu, Yao Qian, Chengyi Wang, Shujie Liu, Yanmin Qian, and Michael Zeng. 2022. Large-scale self-supervised speech representation learning for automatic speaker verification. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6147–6151. IEEE.

Jade Copet, Felix Kreuk, Itai Gat, Tal Remez, David Kant, Gabriel Synnaeve, Yossi Adi, and Alexandre Défossez. 2024. Simple and controllable music generation. *Advances in Neural Information Processing Systems*, 36.

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*.

Brecht Desplanques, Jenthe Thienpondt, and Kris Demuynck. 2020. Ecapa-tdnn: Emphasized channel attention, propagation and aggregation in tdnn based speaker verification. *Interspeech 2020*.

Zhihao Du, Qian Chen, Shiliang Zhang, Kai Hu, Heng Lu, Yexin Yang, Hangrui Hu, Siqi Zheng, Yue Gu, Ziyang Ma, et al. 2024a. Cosyvoice: A scalable multilingual zero-shot text-to-speech synthesizer based on supervised semantic tokens. *arXiv preprint arXiv:2407.05407*.

Zhihao Du, Yuxuan Wang, Qian Chen, Xian Shi, Xiang Lv, Tianyu Zhao, Zhifu Gao, Yexin Yang, Changfeng Gao, Hui Wang, et al. 2024b. Cosyvoice 2: Scalable streaming speech synthesis with large language models. *arXiv preprint arXiv:2412.10117*.

Jesse Engel, Lamtharn Hantrakul, Chenjie Gu, and Adam Roberts. 2020. Ddsp: Differentiable digital signal processing. *arXiv preprint arXiv:2001.04643*.

Zhifu Gao, Zerui Li, Jiaming Wang, Haoneng Luo, Xian Shi, Mengzhe Chen, Yabin Li, Lingyun Zuo, Zhihao Du, Zhangyu Xiao, et al. 2023. Funasr: A fundamental end-to-end speech recognition toolkit. *arXiv preprint arXiv:2305.11013*.

Hugo Flores Garcia, Prem Seetharaman, Rithesh Kumar, and Bryan Pardo. 2023. Vampnet: Music generation via masked acoustic token modeling. *arXiv preprint arXiv:2307.04686*.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2020. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144.

Alexey Gritsenko, Tim Salimans, Rianne van den Berg, Jasper Snoek, and Nal Kalchbrenner. 2020. A spectral energy distance for parallel speech synthesis. *Advances in Neural Information Processing Systems*, 33:13062–13072.

Hao-Han Guo, Kun Liu, Fei-Yu Shen, Yi-Chen Wu, Feng-Long Xie, Kun Xie, and Kai-Tuo Xu. 2024. Fireredtts: A foundation text-to-speech framework for industry-level generative speech applications. *arXiv preprint arXiv:2409.03283*.

Zhifang Guo, Yichong Leng, Yihan Wu, Sheng Zhao, and Xu Tan. 2023. Prompttts: Controllable text-to-speech with text descriptions. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Haorui He, Zengqiang Shang, Chaoren Wang, Xuyuan Li, Yicheng Gu, Hua Hua, Liwei Liu, Chen Yang, Jiaqi Li, Peiyang Shi, et al. 2024. Emilia: An extensive, multilingual, and diverse speech dataset for large-scale speech generation. In *2024 IEEE Spoken Language Technology Workshop (SLT)*, pages 885–890. IEEE.

Zhichao Huang, Chutong Meng, and Tom Ko. 2023. Repcodec: A speech representation codec for speech tokenization. *arXiv preprint arXiv:2309.00169*.

Philip Jackson and SJUoSG Haq. 2014. Surrey audio-visual expressed emotion (savee) database. *University of Surrey: Guildford, UK*.

Jesin James, Li Tian, and Catherine Watson. 2018. An open source emotional speech corpus for human robot interaction applications. *Interspeech 2018*.

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

Shengpeng Ji, Jialong Zuo, Minghui Fang, Ziyue Jiang, Feiyang Chen, Xinyu Duan, Baoxing Huai, and Zhou Zhao. 2024b. Textrolspeech: A text style control speech corpus with codec language text-to-speech models. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 10301–10305. IEEE.

Zeyu Jin, Jia Jia, Qixin Wang, Kehan Li, Shuoyi Zhou, Songtao Zhou, Xiaoyu Qin, and Zhiyong Wu. 2024. Speechcraft: A fine-grained expressive speech dataset with natural language description. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 1255–1264.

Yuma Koizumi, Heiga Zen, Shigeki Karita, Yifan Ding, Kohei Yatabe, Nobuyuki Morioka, Michiel Bacchiani, Yu Zhang, Wei Han, and Ankur Bapna. 2023. Libritts-r: A restored multi-speaker text-to-speech corpus. *arXiv preprint arXiv:2305.18802*.

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.

Felix Kreuk, Gabriel Synnaeve, Adam Polyak, Uriel Singer, Alexandre Défossez, Jade Copet, Devi Parikh, Yaniv Taigman, and Yossi Adi. 2023. Audiogen: Textually guided audio generation. In *The Eleventh International Conference on Learning Representations*.

Kundan Kumar, Rithesh Kumar, Thibault De Boissiere, Lucas Gestin, Wei Zhen Teoh, Jose Sotelo, Alexandre De Brebisson, Yoshua Bengio, and Aaron C Courville. 2019. Melgan: Generative adversarial networks for conditional waveform synthesis. *Advances in neural information processing systems*, 32.

Rithesh Kumar, Prem Seetharaman, Alejandro Luebs, Ishaan Kumar, and Kundan Kumar. 2024. High-fidelity audio compression with improved rvqgan. *Advances in Neural Information Processing Systems*, 36.

Gael Le Lan, Varun Nagaraja, Ernie Chang, David Kant, Zhaoheng Ni, Yangyang Shi, Forrest Iandola, and Vikas Chandra. 2024. Stack-and-delay: a new codebook pattern for music generation. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 796–800. IEEE.

Keon Lee, Kyumin Park, and Daeyoung Kim. 2023. Dailytalk: Spoken dialogue dataset for conversational text-to-speech. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Hanzhao Li, Liumeng Xue, Haohan Guo, Xinfa Zhu, Yuanjun Lv, Lei Xie, Yunlin Chen, Hao Yin, and Zhifei Li. 2024a. Single-codec: Single-codebook speech codec towards high-performance speech generation. *arXiv preprint arXiv:2406.07422*.

Xu Li, Qirui Wang, and Xiaoyu Liu. 2024b. Masksr: Masked language model for full-band speech restoration. *arXiv preprint arXiv:2406.02092*.

Yue Li, Xinsheng Wang, Li Zhang, and Lei Xie. 2024c. Scdnet: Self-supervised learning feature-based speaker change detection. *arXiv preprint arXiv:2406.08393*.

Zheng Lian, Haiyang Sun, Licai Sun, Kang Chen, Mngyu Xu, Kexin Wang, Ke Xu, Yu He, Ying Li, Jinming Zhao, et al. 2023. Mer 2023: Multi-label learning, modality robustness, and semi-supervised learning. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 9610–9614.

Rui Liu, Yifan Hu, Yi Ren, Xiang Yin, and Haizhou Li. 2024. Generative expressive conversational speech synthesis. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 4187–4196.

Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. 2022. A convnet for the 2020s. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11976–11986.

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

Dan Lyth and Simon King. 2024. Natural language guidance of high-fidelity text-to-speech with synthetic annotations. *arXiv preprint arXiv:2402.01912*.

Linhan Ma, Dake Guo, Kun Song, Yuepeng Jiang, Shuai Wang, Liumeng Xue, Weiming Xu, Huan Zhao, Binbin Zhang, and Lei Xie. 2024. Wenetspeech4tts: A 12,800-hour mandarin tts corpus for large speech generation model benchmark. *arXiv preprint arXiv:2406.05763*.

Ziyang Ma, Zhisheng Zheng, Jiaxin Ye, Jinchao Li, Zhifu Gao, Shiliang Zhang, and Xie Chen. 2023. emotion2vec: Self-supervised pre-training for speech emotion representation. *arXiv preprint arXiv:2312.15185*.

MagicData. 2019. Magicdata mandarin chinese read speech corpus.

Luz Martinez, Mohammed Abdelwahab, and Carlos Busso. 2020. The msp-conversation corpus. *Interspeech 2020*.

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

Tu Anh Nguyen, Wei-Ning Hsu, Antony d'Avirro, Bowen Shi, Itai Gat, Maryam Fazel-Zarani, Tal Remez, Jade Copet, Gabriel Synnaeve, Michael Hassid, et al. 2023. Expresso: A benchmark and analysis of discrete expressive speech resynthesis. *arXiv preprint arXiv:2308.05725*.

Kari Ali Noriy, Xiaosong Yang, and Jian Jun Zhang. 2023. Emns/imz/corpus: An emotive single-speaker dataset for narrative storytelling in games, television and graphic novels. *arXiv preprint arXiv:2305.13137*.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. Librispeech: an asr corpus based on public domain audio books. In *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 5206–5210. IEEE.

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*.

Ankita Pasad, Bowen Shi, and Karen Livescu. 2023. Comparative layer-wise analysis of self-supervised speech models. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. 2018. Meld: A multimodal multi-party dataset for emotion recognition in conversations. *arXiv preprint arXiv:1810.02508*.

Vineel Pratap, Qiantong Xu, Anuroop Sriram, Gabriel Synnaeve, and Ronan Collobert. 2020. Mls: A large-scale multilingual dataset for speech research. *ArXiv*, abs/2012.03411.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*, pages 28492–28518. PMLR.

Yong Ren, Tao Wang, Jiangyan Yi, Le Xu, Jianhua Tao, Chu Yuan Zhang, and Junzuo Zhou. 2024. Fewer-token neural speech codec with time-invariant codes. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 12737–12741. IEEE.

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

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

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*.

Dehua Tao, Daxin Tan, Yu Ting Yeung, Xiao Chen, and Tan Lee. 2024. Toneunit: A speech discretization approach for tonal language speech synthesis. *arXiv preprint arXiv:2406.08989*.

Jianhua Tao, Fangzhou Liu, Meng Zhang, and Huibin Jia. 2008. Design of speech corpus for mandarin text to speech. In *The blizzard challenge 2008 workshop*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

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

Apoorv Vyas, Bowen Shi, Matthew Le, Andros Tjandra, Yi-Chiao Wu, Baishan Guo, Jiemin Zhang, Xinyue Zhang, Robert Adkins, William Ngan, et al. 2023. Audiobox: Unified audio generation with natural language prompts. *arXiv preprint arXiv:2312.15821*.

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

Hongji Wang, Chengdong Liang, Shuai Wang, Zhengyang Chen, Binbin Zhang, Xu Xiang, Yanlei Deng, and Yanmin Qian. 2023b. Wespeaker: A research and production oriented speaker embedding learning toolkit. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Kaisiyuan Wang, Qianyi Wu, Linsen Song, Zhuoqian Yang, Wayne Wu, Chen Qian, Ran He, Yu Qiao, and Chen Change Loy. 2020. Mead: A large-scale audio-visual dataset for emotional talking-face generation. In *European Conference on Computer Vision*, pages 700–717. Springer.

Xiaofei Wang, Manthan Thakker, Zhuo Chen, Naoyuki Kanda, Sefik Emre Eskimez, Sanyuan Chen, Min Tang, Shujie Liu, Jinyu Li, and Takuya Yoshioka. 2024a. Speechx: Neural codec language model as a versatile speech transformer. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*.

Yuancheng Wang, Haoyue Zhan, Liwei Liu, Ruihong Zeng, Haotian Guo, Jiachen Zheng, Qiang Zhang, Xueyao Zhang, Shunsi Zhang, and Zhizheng Wu. 2024b. Maskgct: Zero-shot text-to-speech with masked generative codec transformer. *arXiv preprint arXiv:2409.00750*.

Zhichao Wang, Yuanzhe Chen, Xinsheng Wang, Lei Xie, and Yuping Wang. 2024c. Streamvoice+: Evolving into end-to-end streaming zero-shot voice conversion. *IEEE Signal Processing Letters*.

Haibin Wu, Naoyuki Kanda, Sefik Emre Eskimez, and Jinyu Li. 2024. Ts3-codec: Transformer-based simple streaming single codec. *arXiv preprint arXiv:2411.18803*.

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

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

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. 2024. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*.

Zhen Ye, Zeqian Ju, Haohe Liu, Xu Tan, Jianyi Chen, Yiwen Lu, Peiwen Sun, Jiahao Pan, Weizhen Bian, Shulin He, et al. 2024a. Flashspeech: Efficient zero-shot speech synthesis. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 6998–7007.

Zhen Ye, Peiwen Sun, Jiahe Lei, Hongzhan Lin, Xu Tan, Zheqi Dai, Qiuqiang Kong, Jianyi Chen, Jiahao Pan, Qifeng Liu, et al. 2024b. Codec does matter: Exploring the semantic shortcoming of codec for audio language model. *arXiv preprint arXiv:2408.17175*.

Zhen Ye, Xinfa Zhu, Chi-Min Chan, Xinsheng Wang, Xu Tan, Jiahe Lei, Yi Peng, Haohe Liu, Yizhu Jin, Zheqi DAI, et al. 2025. Llasa: Scaling train-time and inference-time compute for llama-based speech synthesis. *arXiv preprint arXiv:2502.04128*.

Jiahui Yu, Xin Li, Jing Yu Koh, Han Zhang, Ruoming Pang, James Qin, Alexander Ku, Yuanzhong Xu, Jason Baldridge, and Yonghui Wu. 2021. Vector-quantized image modeling with improved vqgan. *arXiv preprint arXiv:2110.04627*.

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*.

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*.

Jinming Zhao, Tenggan Zhang, Jingwen Hu, Yuchen Liu, Qin Jin, Xinchao Wang, and Haizhou Li. 2022. M3ed: Multi-modal multi-scene multi-label emotional dialogue database. *arXiv preprint arXiv:2205.10237*.

Youqiang Zheng, Weiping Tu, Yueteng Kang, Jie Chen, Yike Zhang, Li Xiao, Yuhong Yang, and Long Ma. 2024. Freecodec: A disentangled neural speech codec with fewer tokens. *arXiv preprint arXiv:2412.01053*.

Kun Zhou, Berrak Sisman, Rui Liu, and Haizhou Li. 2021. Seen and unseen emotional style transfer for voice conversion with a new emotional speech dataset. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 920–924. IEEE.

Shuoyi Zhou, Yixuan Zhou, Weiqing Li, Jun Chen, Runchuan Ye, Weihao Wu, Zijian Lin, Shun Lei, and Zhiyong Wu. 2024a. The codec language model-based zero-shot spontaneous style tts system for covoc challenge 2024. In *2024 IEEE 14th International Symposium on Chinese Spoken Language Processing (ISCSLP)*, pages 496–500. IEEE.

Yixuan Zhou, Xiaoyu Qin, Zeyu Jin, Shuoyi Zhou, Shun Lei, Songtao Zhou, Zhiyong Wu, and Jia Jia. 2024b. Voxinstruct: Expressive human instruction-to-speech generation with unified multilingual codec language modelling. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 554–563.

Xinfa Zhu, Wenjie Tian, Xinsheng Wang, Lei He, Yujia Xiao, Xi Wang, Xu Tan, Sheng Zhao, and Lei Xie. 2024. Unistyle: Unified style modeling for speaking style captioning and stylistic speech synthesis. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 7513–7522.

Alon Ziv, Itai Gat, Gael Le Lan, Tal Remez, Felix Kreuk, Alexandre Défossez, Jade Copet, Gabriel Synnaeve, and Yossi Adi. 2024. Masked audio generation using a single non-autoregressive transformer. *arXiv preprint arXiv:2401.04577*.

A BiCodec

The model structure of BiCodec is illustrated in Fig. 6. BiCodec primarily consists of three components:

- Semantic Tokenizer
- Global Tokenizer
- Decoder

Additionally, to compute the feature loss with the input wav2vec 2.0 features, an extra ConvNeXt block is incorporated to predict wav2vec 2.0 features, to further ensure the semantic relevance.

A.1 Model Configurations

The semantic tokenizer consists of 12 ConvNeXt blocks and 2 downsampling blocks. The downsampling blacks is only for semantic codes with lower than 50 TPS. The codebook size of VQ is 8192. The ECAPA-TDNN in the global tokenizer features an embedding dimension of 512. Meanwhile, the vector number of the learnable queries in the global tokenizer equal to the final goal token sequence length. For the FSQ module, the FSQ dimension is set to 6, with each dimension having 4 levels, resulting in a codebook size of 4096.

The upsampling rates in the Transposed Convolution Blocks are set to [8, 5, 4, 2] for 16 kHz sampled audio and [8, 5, 4, 3] for 24 kHz sampled audio. The reconstruction performance of BiCodec with 24 kHz sampled audio is presented in Table 9.

A.2 Compared Methods

- **Encodec** (Défossez et al., 2022): An RVQ-based codec designed for universal audio compression.
- **DAC** (Kumar et al., 2024): An RVQ-based codec for universal audio.
- **Mimi** (Défossez et al., 2024): An RVQ-based codec with semantic constraint for speech.

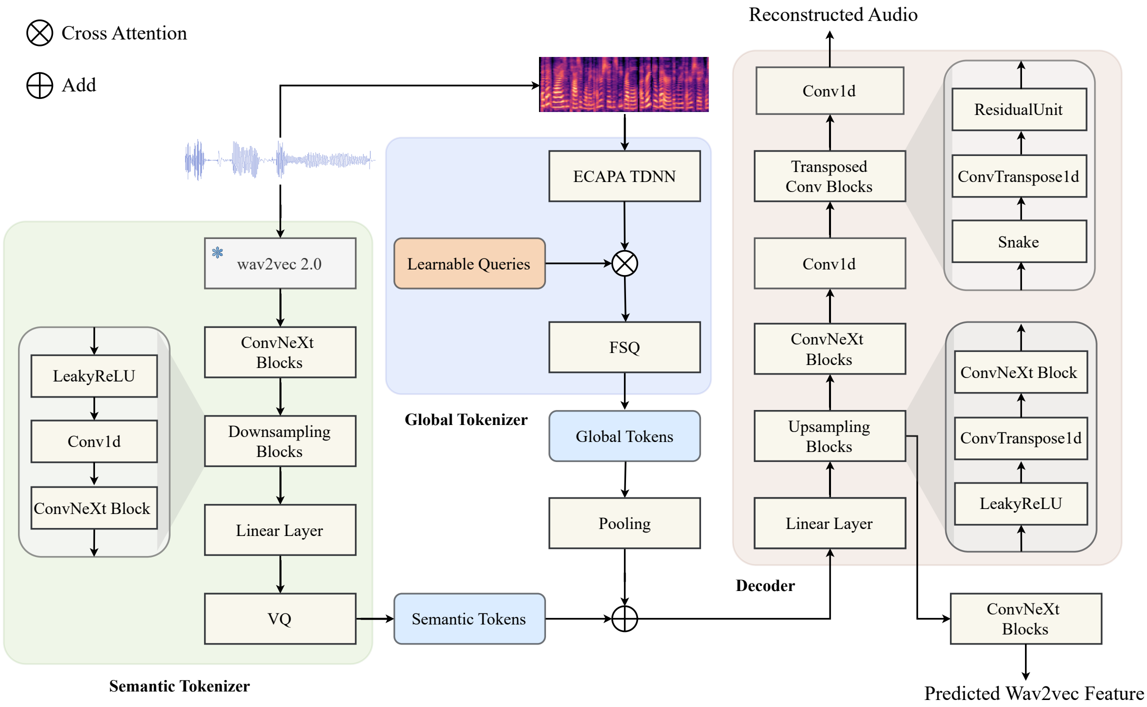


Figure 6: Model Structure of BiCodec

- **Single-Codec** (Li et al., 2024a): A single-stream Mel codec that incorporates speaker embeddings. The reconstruction results for this method are provided by the authors.
- **BigCodec** (Xin et al., 2024): A VQ-based single-stream codec for speech.
- **SpeechTokenizer** (Zhang et al., 2023a): An RVQ-based codec with semantic distillation for speech.
- **X-codec** (Ye et al., 2024b): An RVQ-based codec with semantic distillation for speech.
- **X-codec2** (Ye et al., 2025): A FSQ-based single-stream codec with semantic distillation for speech.
- **StableCodec** (Parker et al., 2024): A residual FSQ-based tokenizer for speech.
- **WavTokenizer** (Ji et al., 2024a): A single VQ codebook-based tokenizer for universal audio.

A.3 Additional Experiment

To evaluate the performance of BiCodec at lower bitrates, we apply a downsampling operation in the semantic encoder, reducing the semantic token rate to 25 TPS. We compare BiCodec with Single-Codec (Li et al., 2024a), which operates at a similar bitrate, on the LibriSpeech test-clean and LibriTTS test-clean datasets. The results are presented in Table 6 and Table 7.

Global Token Length The reconstruction performance of BiCodec with varying global token lengths on the LibriTTS test-clean dataset is presented in Table 8.

Performance on Other Datasets To evaluate the generalization ability of BiCodec, we conducted experiments on a broader range of diverse datasets. The results are presented in Table 9.

B Inference of Spark-TTS

Zero-shot TTS There are two inference strategies for zero-shot TTS:

- Using the text to be synthesized along with the global tokens from a reference audio as the prompt to generate speech, e.g., [*<content text>* *<global token>* $\rightarrow$ *<semantic token>*].
- Incorporating both the transcript and semantic tokens of the reference audio as a prefix in the prompt, e.g., [*<content text>* *<reference text>* *<global token>* *<semantic token of reference>* $\rightarrow$ *<semantic token>*].

Among these, the second approach achieves higher speaker similarity. The results reported in Table 4 are based on this second inference strategy. A comparison between the two inference methods is provided in Table 10.

Table 6: Performance of BiCodec with lower bitrate on the LibriSpeech test-clean dataset.

Model	Codebook Size	Nq	Token Rate	Bandwidth	STOI↑	PESQ NB↑	PESQ WB↑	UTMOS↑	SIM↑
Single-Codec	8192	1	23.4	304	0.86	2.42	1.88	3.72	0.60
BiCodec-4096-25	4096	1	25	300	0.88	2.53	1.97	4.00	0.70
BiCodec-8192-25	8192	1	25	325	0.89	2.62	2.05	4.13	0.71
BiCodec-4096-50	4096	1	50	600	0.92	3.03	2.42	4.17	0.78

Table 7: Reconstruction performance of BiCodec with various bitrates on the LibriTTS test-clean dataset.

Codebook Size	Nq	Token Rate	Bandwidth	STOI↑	PESQ NB↑	PESQ WB↑	UTMOS↑	SIM↑
4096	1	25	300	0.88	2.47	1.91	3.88	0.67
8192	1	25	325	0.88	2.56	1.98	4.02	0.68
4096	1	50	600	0.91	2.96	2.36	4.10	0.75
8192	1	50	650	**0.92**	**3.08**	**2.46**	**4.11**	**0.78**

Table 8: Performance of BiCodec with varying global token lengths for reconstruction on the LibriTTS test-clean dataset, where "w/o" indicates the omission of FSQ-based quantization, and gvq-32 means the global tokenizer is implemented with group VQ.

Global Token	STOI↑	PESQ NB↑	PESQ WB↑	UTMOS↑	SIM↑
w/o	**0.923**	**3.1**	**2.48**	4.09	**0.81**
gvq-32	0.913	2.91	2.30	4.06	0.71
8	0.916	2.97	2.34	4.10	0.72
16	0.918	3.03	2.40	4.08	0.74
32	0.921	**3.08**	**2.46**	**4.11**	0.78

Voice Creation Controllable TTS includes two levels of control for inference:

- Coarse-grained control: The prompt consists of the text to be synthesized along with attribute labels, e.g., [<*content text*> <*attribute label*> $\rightarrow$ <*attribute values*> <*global tokens*> <*semantic token*>]. In this process, the fine-grained attribute values are predicted first, followed by the generation of global tokens and then semantic tokens, in a CoT manner.

- Fine-grained control: The prompt includes the text to be synthesized, attribute levels, and precise attribute values, e.g., [<*content text*> <*attribute label*> <*attribute values*> $\rightarrow$ <*global tokens*> <*semantic token*>].

C Compared Zero-shot Methods

- **Seed-TTS** (Anastassiou et al., 2024): A two-stage model that employs an AR LM for semantic token prediction and flow matching for acoustic feature generation.

- **FireRedTTS** (Guo et al., 2024): A two-stage model similar to Seed-TTS, using an AR LM for semantic tokens and flow matching for acoustic features.

- **MaskGCT** (Wang et al., 2024b): A NAR model that applies masking-based generative strategies for speech synthesis.

- **E2 TTS** : A flow matching-based model that predicts Mel spectrograms as acoustic features.

- **F5-TTS** (Chen et al., 2024b): A flow matching-based method that also uses Mel spectrograms as acoustic features.

- **CosyVoice** (Du et al., 2024a): A two-stage model with an AR LM for semantic token prediction and flow matching for acoustic feature generation.

- **CosyVoice2** (Du et al., 2024b): An improved version of CosyVoice, maintaining the two-stage structure with an AR LM for semantic tokens and flow matching for acoustic features.

Table 9: Reconstruction performance on various datasets: Data-P comprises low-quality Chinese recordings made by internal staff using mobile phones; Data-S consists of expressive Chinese data recorded in a professional studio; and Data-M is a multilingual dataset collected from in-the-wild sources.

Data	Method	Codebook Size	Traing Data	STOI↑	PESQ NB↑	PESQ WB↑	UTMOS↑	SIM↑
Data-P	X-codec2	65536	150k	0.89	2.69	2.10	3.16	0.73
	BiCodec	8192	3k	**0.90**	**2.80**	**2.22**	**3.22**	**0.78**
	BiCodec-24k	8192	20k	**0.90**	**2.80**	2.19	3.20	**0.78**
Data-S	X-codec2	65536	150k	0.92	2.81	2.30	3.16	0.69
	BiCodec	8192	3k	0.93	**3.04**	**2.50**	**3.28**	**0.82**
	BiCodec-24k	8192	20k	0.93	3.00	2.44	3.24	**0.82**
Data-M	X-codec2	65536	150k	0.84	2.43	1.87	2.17	0.75
	BiCodec	8192	3k	**0.85**	2.56	**1.91**	2.17	**0.76**
	BiCodec-24k	8192	20k	**0.85**	**2.57**	**1.91**	**2.28**	**0.76**

- **Llasa** (Ye et al., 2025): A single-stream codec-based TTS model that uses a single AR language model for direct single-stream code prediction.

Table 10: Zero-shot performance of Spark-TTS with and without reference audio as a prefix.

Model	test-zh		test-en	
	CER↓	SIM↑	WER↓	SIM↑
Spark-TTS	1.20	0.678	1.98	0.584
Spark-TTS w/o prefix	**0.98**	0.628	**1.32**	0.474

D Objective Metircs

- **STOI** (Andersen et al., 2017): A widely used metric for assessing speech intelligibility. Scores range from 0 to 1, with higher values indicating better intelligibility.
- **PESQ** (Rix et al., 2001): A speech quality assessment metric that compares the reconstructed speech to a reference speech signal. We evaluate using both wide-band (WB) and narrow-band (NB) settings.
- **UTMOS** (Saeki et al., 2022): An automatic Mean Opinion Score (MOS) predictor, providing an estimate of overall speech quality.
- **SIM**: A speaker similarity metric, computed as the cosine similarity between the speaker embeddings of the reconstructed speech (generated speech in TTS) and the original input speech (prompt speech in TTS). We extract speaker embeddings using WavLM-large, fine-tuned on the speaker verification task (Chen et al., 2022).

E VoxBox

E.1 Criteria for Pitch and Speed Categorization

- **Speed** The adoption of the 5th, 20th, 80th, and 95th percentiles to segment speech rates into distinct categories is founded on the need to accurately reflect the natural distribution of speech tempo variations within the population. These percentiles help to capture the extremes and the more central values of speech rate, ensuring that each category is meaningful and representative of specific vocal characteristics.

- **Pitch** Similar to the segmentation of speech rate, the division of pitch also starts from human subjective perception and the actual distribution characteristics. However, because humans are more sensitive to higher frequencies within the range of human fundamental frequencies, the 5th, 20th, 70th, and 90th percentiles are used as the division boundaries.

Pitch Group for Male

Very Low:	$<$ 145 Mel
Low:	145–164 Mel
Moderate:	164–211 Mel
High:	211–250 Mel
Very High:	$>=$ 250 Mel

Pitch Group for Female

Very Low: < 225 Mel
Low: 225–258 Mel
Moderate: 258–314 Mel
High: 314–353 Mel
Very High: >= 353 Mel

Speaking Rate Group for Chinese

Very Slow: < 2.7 SPS
Slow: 2.7–3.6 SPS
Moderate: 3.6–5.2 SPS
Fast: 5.2–6.1 SPS
Very Fast: >= 6.1 SPS

Speaking Rate Group for English

Very Slow: < 2.6 SPS
Slow: 2.6–3.4 SPS
Moderate: 3.4–4.8 SPS
Fast: 4.8–5.5 SPS
Very Fast: >= 5.5 SPS

E.2 Data for Gender Predictor Training

We fine-tune the WavLM-large model for gender classification using datasets that contain explicit gender labels, including VCTK (Yamagishi et al., 2019), AISHELL-3 (Shi et al., 2020), MLS-English (Pratap et al., 2020), MAGICDATA (MagicData, 2019), and CommonVoice (Ardila et al., 2019).

E.3 Annotation

In addition to the attributes involved in the experiments of this paper, to make VoxBox applicable to a wider range of scenarios, we have also annotated more information for each sample of VoxBox, including age and emotion. Similar to the gender annotations, we fine-tune the WavLM-large model based on AISHELL-3, VCTK, MAGICDATA, CommonVoice, and HQ-Conversations to predict five age ranges: Child, Teenager, Young Adult, Middle-aged, and Elderly. The performance metrics for both the gender and age predictors are presented in Table 11, where both Wav2vec 2.0-ft (Burkhardt et al., 2023) and SpeechCraft (Jin et al., 2024) are based on the pre-trained Wav2vec 2.0 model.

For datasets without emotion labels in the original metadata, we assign various emotion labels, sourced from different models, to the relevant samples. Specifically, we provide the following tags:

Table 11: Comparison of different models on attribute predictions: All evaluations are conducted on the AISHELL-3 test dataset.

Model	Age Acc↑	Gender Acc↑
wav2vec 2.0-ft	80.2	98.8
SpeechCraft	87.7	97.7
Our	**95.6**	**99.4**

- **emotion2vec Emotion**: Emotion label predicted with Emtion2vec (Ma et al., 2023).
- **Confidence Score**: Confidence score of the the predicted emotion2vec label given by emotion2vec.
- **SenseVoiceSmall Emotion**: Emotion label predicted with SenseVoiceSmall[8].
- **Text Emotion**: Emotion label predicted with Qwen2.5-72B-Instruct [9] with text as input. The prompt case for English text can be found in Box

Prompt for Text Emotion Tag (English)

Please assess the emotion of the following text and select the most appropriate label from these options:
[Fearful, Happy, Disgusted, Sad, Surprised, Angry, Neutral].
Please note, only provide the label without any additional description or reasoning.
Here is the text: "Clearly, the need for a personal loan is written in the stars."

E.4 Data Statistics

The distributions of speaking rate, duration, and pitch are shown in Fig 7, while the distributions of gender and age are presented in Fig 8.

E.5 Source Data

- **AISHELL-3**: A multi-speaker Mandarin speech corpus for TTS. Source: https://www.openslr.org/93/

[8]https://huggingface.co/FunAudioLLM/SenseVoiceSmall

[9]https://huggingface.co/Qwen/Qwen2.5-72B-Instruct

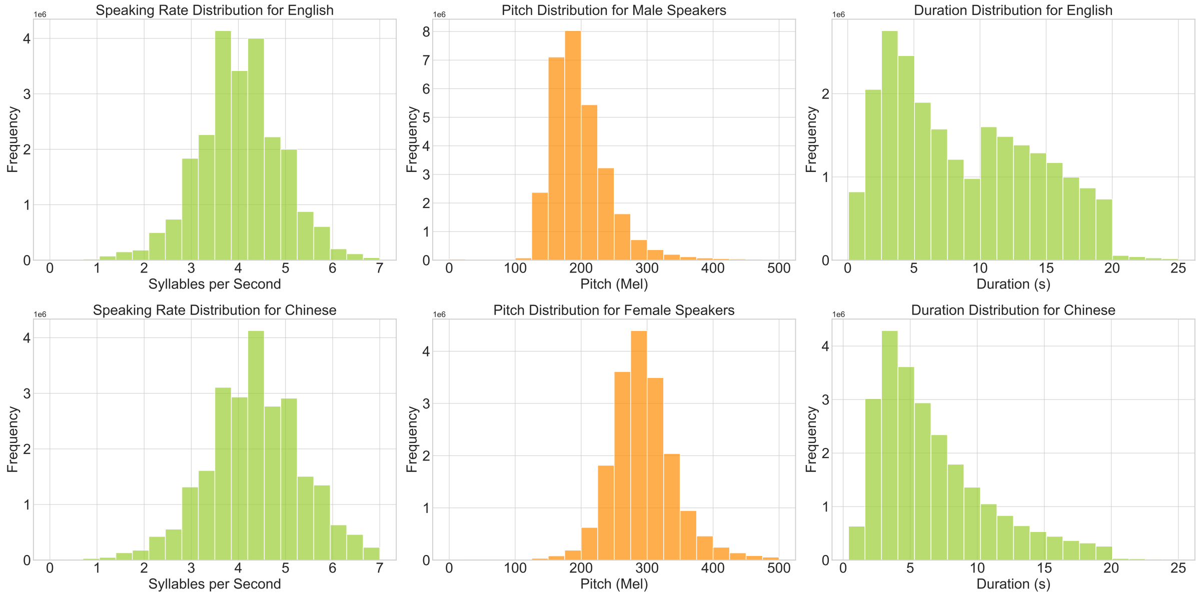


Figure 7: Data distribution of VoxBox.

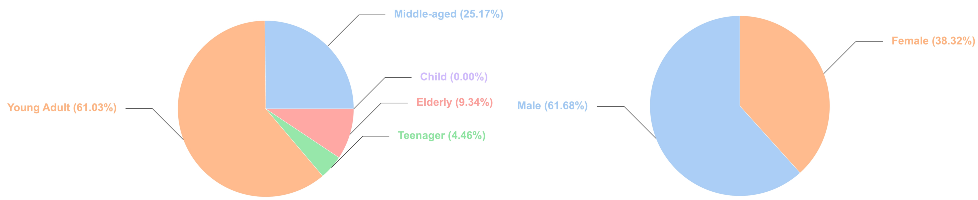


Figure 8: Gender and age distribution of VoxBox.

- **CASIA**: An emotional multi-speaker Mandarin speech corpus containing six emotions for TTS. Source: https://gitcode.com/open-source-toolkit/bc5e6
- **CREMA-D**: An emotional multi-speaker multilingual speech corpus containing six emotions and four intensity levels for TTS. Source:https://github.com/CheyneyComputerScience/CREMA-D
- **Dailytalk**: A multi-speaker English speech corpus with conversational style for TTS. Source:https://github.com/keonlee9420/DailyTalk
- **Emilia**: A multi-speaker multilingual speech corpus containing six languages for TTS. Source: https://emilia-dataset.github.io/Emilia-Demo-Page/
- **EMNS**: An emotional single-speaker English speech corpus for TTS. Source: https://www.openslr.org/136
- **EmoV-DB**: An emotional multi-speaker English speech corpus containing four emotions for TTS. Source: https://mega.nz/folder/KBp32apT#gLIgyWf9iQ-yqnWFUFuUHg/mYwUnI4K
- **ESD**: An emotional multi-speaker bilingual speech corpus containing five emotions for TTS. Source: https://hltsingapore.github.io/ESD/
- **Expresso**: A multi-speaker English speech corpus with reading and improvising conversational style for TTS. Source: https://speechbot.github.io/expresso/
- **Gigaspeech**: A multi-speaker English speech corpus with reading style for TTS. Source: https://github.com/SpeechColab/GigaSpeech
- **Hi-Fi TTS**: A multi-speaker English speech corpus with reading style for TTS. Source: https://openslr.org/109/
- **HQ-Conversations**: A mutli-speaker Mandarin speech corpus with conversational style for TTS. Source: https://www.magicdatatech.com/iscslp-2024/

- **IEMOCAP**: An emotional multi-speaker English speech corpus containing five emotions for TTS. Source: https://sail.usc.edu/iemocap/iemocap_release.htm

- **JL-Corpus**: An emotional multi-speaker English speech corpus containing five primary emotions and five secondary emotions for TTS. Source: https://www.kaggle.com/datasets/tli725/jl-corpus

- **Librispeech**: A mutli-speaker English speech corpus with reading style for TTS. Source: https://tensorflow.google.cn/datasets/catalog/librispeech

- **LibriTTS-R**: Sound quality improved version of the LibriTTS (Zen et al., 2019) corpus which is a large-scale corpus of English speech for TTS. Source: https://www.openslr.org/141/

- **M3ED**: An emotional mutli-speaker Mandarin speech corpus containing seven emotions for TTS. Source: https://github.com/aim3-ruc/rucm3ed

- **MAGICDATA**: A mutli-speaker Mandarin speech corpus with conversational style for TTS. Source: https://openslr.org/68/

- **MEAD**: An emotional mutli-speaker English speech corpus containing eight emotions and three intensity levels for TTS. Source: https://github.com/uniBruce/Mead

- **MELD**: An emotional mutli-speaker English speech corpus containing seven emotions for TTS. Source: https://affective-meld.github.io/

- **MER2023**: An emotional mutli-speaker Mandarin speech corpus containing six emotions for TTS. Source: http://www.merchallenge.cn/datasets

- **MLS-English**: A mutli-speaker English speech corpus for TTS. Source: https://www.openslr.org/94/

- **MSP-Podcast**: An emotional mutli-speaker English speech corpus containing eight emotions for TTS. Source: https://ecs.utdallas.edu/research/researchlabs/msp-lab/MSP-Podcast.html

- **NCSSD-CL**: A mutli-speaker bilingual speech corpus for TTS. Source: https://github.com/uniBruce/Mead

- **NCSSD-RL**: A mutli-speaker bilingual speech corpus for TTS. Source: https://github.com/uniBruce/Mead

- **RAVDESS**: An emotional mutli-speaker English speech corpus containing eight emotions and two intensity levels for TTS. Source: https://www.kaggle.com/datasets/uwrfkaggler/ravdess-emotional-speech-audio

- **SAVEE**: An emotional mutli-speaker English speech corpus containing seven emotions for TTS. Source: https://www.kaggle.com/datasets/ejlok1/surrey-audiovisual-expressed-emotion-savee

- **TESS**: An emotional mutli-speaker English speech corpus containing seven emotions for TTS. Source: https://tspace.library.utoronto.ca/handle/1807/24487

- **VCTK**: A mutli-speaker English speech corpus for TTS. Source: https://datashare.ed.ac.uk/handle/10283/2651

- **WenetSpeech4TTS**: A large-scale mutli-speaker Mandarin speech corpus for TTS. Source: https://wenetspeech4tts.github.io/wenetspeech4tts/

F SparkVox: A Toolkit for Speech Related Tasks

The training code for Spark-TTS will be integrated into the open-source SparkVox framework.SparkVox is a training framework designed for speech-related tasks, supporting a variety of applications, including: vocoder, codec, TTS, and speech understanding. Additionally, SparkVox provides various file processing tools for both text and speech data, facilitating efficient data handling. Its simplified framework structure is illustrated in Fig. 9.

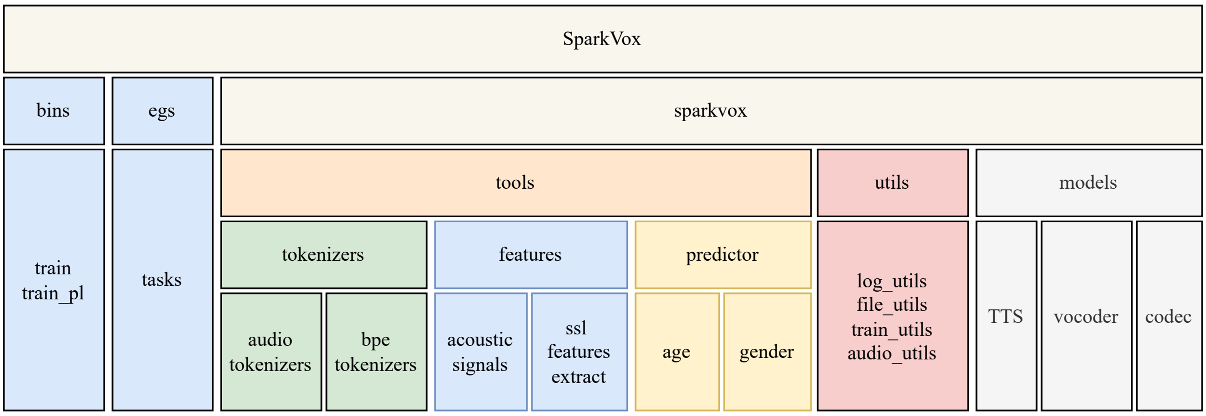


Figure 9: Framework of SparkVox.

Table 12: VoxBox Statistics

Data	Language	#Utterance	Duration (h)		
			Male	Female	Total
AISHELL-3 (Shi et al., 2020)	Chinese	88,035	16.01	69.61	85.62
CASIA (Tao et al., 2008)	Chinese	857	0.25	0.2	0.44
Emilia-CN (He et al., 2024)	Chinese	15,629,241	22,017.56	12,741.89	34,759.45
ESD (Zhou et al., 2021)	Chinese	16,101	6.69	7.68	14.37
HQ-Conversations (Zhou et al., 2024a)	Chinese	50,982	35.77	64.23	100
M3ED (Zhao et al., 2022)	Chinese	253	0.04	0.06	0.1
MAGICDATA (MagicData, 2019)	Chinese	609,474	360.31	393.81	754.13
MER2023 (Lian et al., 2023)	Chinese	1,667	0.86	1.07	1.93
NCSSD-CL-CN (Liu et al., 2024)	Chinese	98,628	53.83	59.21	113.04
NCSSD-RC-CN (Liu et al., 2024)	Chinese	21,688	7.05	22.53	29.58
WenetSpeech4TTS (Ma et al., 2024)	Chinese	8,856,480	7,504.19	4,264.3	11,768.49
Total	Chinese	25,373,406	30,002.56	17,624.59	47,627.15
CREMA-D (Cao et al., 2014)	English	809	0.3	0.27	0.57
Dailytalk (Lee et al., 2023)	English	23,754	10.79	10.86	21.65
EmiliaEN (He et al., 2024)	English	8,303,103	13,724.76	6,573.22	20,297.98
EMNS (Noriy et al., 2023)	English	918	0	1.49	1.49
EmoV-DB (Adigwe et al., 2018)	English	3,647	2.22	2.79	5
Expresso (Nguyen et al., 2023)	English	11,595	5.47	5.39	10.86
Gigaspeech (Chen et al., 2021)	English	6,619,339	4,310.19	2,885.66	7,195.85
Hi-Fi TTS (Bakhturina et al., 2021)	English	323,911	133.31	158.38	291.68
IEMOCAP (Busso et al., 2008)	English	2,423	1.66	1.31	2.97
JL-Corpus (James et al., 2018)	English	893	0.26	0.26	0.52
Librispeech (Panayotov et al., 2015)	English	230,865	393.95	367.67	761.62
LibriTTS-R (Koizumi et al., 2023)	English	363,270	277.87	283.03	560.9
MEAD (Wang et al., 2020)	English	3,767	2.26	2.42	4.68
MELD (Poria et al., 2018)	English	5,100	2.14	1.94	4.09
MLS-English (Pratap et al., 2020)	English	6,319,002	14,366.25	11,212.92	25,579.18
MSP-Podcast (Martinez et al., 2020)	English	796	0.76	0.56	1.32
NCSSD-CL-EN (Liu et al., 2024)	English	62,107	36.84	32.93	69.77
NCSSD-RL-EN (Liu et al., 2024)	English	10,032	4.18	14.92	19.09
RAVDESS (Livingstone and Russo, 2018)	English	950	0.49	0.48	0.97
SAVEE (Jackson and Haq, 2014)	English	286	0.15	0.15	0.31
TESS (Yu et al., 2021)	English	1,956	0	1.15	1.15
VCTK (Yamagishi et al., 2019)	English	44,283	16.95	24.51	41.46
Total	English	22,332,806	33,290.8	21,582.31	54,873.11
Overall Total		47,706,212	63,293.36	39,206.9	102,500.26