What Makes a Good Speech Tokenizer for LLM-Centric Speech Generation? A Systematic Study

Xiaoran Fan[1]*, Zhichao Sun[1]*, Yangfan Gao[1]*, Jingfei Xiong[1]*, Hang Yan[2]*, Yifei Cao[1], Jiajun Sun[1], Shuo Li[1], Zhihao Zhang[1], Zhiheng Xi[1], Yuhao Zhou[1], Senjie Jin[1], Changhao Jiang[1], Junjie Ye[1], Ming Zhang[1], Rui Zheng[2], Zhenhua Han[2], Yunke Zhang[3], Demei Yan[3], Shaokang Dong[3], Tao Ji[1]†, Tao Gui[1,4]†

[1]Fudan University
[2]Shanghai Qiji Zhifeng Co., Ltd
[3]Honor Device Co., Ltd
[4]Shanghai Innovation Institute
{xrfan23, taoji, tgui}@fudan.edu.cn

Abstract

Speech-language models (SLMs) offer a promising path toward unifying speech and text understanding and generation. However, challenges remain in achieving effective cross-modal alignment and high-quality speech generation. In this work, we systematically investigate the role of speech tokenizer designs in LLM-centric SLMs, augmented by speech heads and speaker modeling. We compare coupled, semi-decoupled, and fully decoupled speech tokenizers under a fair SLM framework and find that decoupled tokenization significantly improves alignment and synthesis quality. To address the information density mismatch between speech and text, we introduce multi-token prediction (MTP) into SLMs, enabling each hidden state to decode multiple speech tokens. This results in up to 12× faster decoding and a substantial reduction in word error rate (from 6.07 to 3.01). Furthermore, we propose a speaker-aware generation paradigm and introduce RoleTriviaQA, a large-scale role-playing knowledge QA benchmark with diverse speaker identities. Experiments demonstrate that our methods enhance both knowledge understanding and speaker consistency.

Github —
https://github.com/cnxupupup/SLM-Decoupled-MTP
Extended version — https://arxiv.org/abs/2506.12537

1 Introduction

In recent years, large language models (LLMs) have achieved remarkable progress in text understanding and generation (Grattafiori et al. 2024; Qwen 2025), fueling rapid advancements in multimodal models such as vision-language models (Liu et al. 2023), vision-language-action models (Kim et al. 2024), and speech-language models (SLMs) (Chu et al. 2023; Li et al. 2025).

Among them, SLMs are particularly valuable for applications like natural and fluent human-computer dialogue (Défossez et al. 2024; Xu et al. 2025) and personalized speech generation (Du et al. 2024a,b). The core challenge lies in generalizing the language understanding and generation capabilities of LLMs to incorporate and integrate speech understanding and generation.

Current approaches for connecting text and speech in SLMs can be broadly categorized into two perspectives. *Speech-centric* models (Du et al. 2024a,b; Anastassiou et al. 2024) are built upon mature speech generation architectures, with LLMs introduced as a conditional context. However, while this approach achieves high-quality speech synthesis, it struggles to fully leverage LLMs' rich world knowledge and powerful capabilities. *LLM-centric* models (Défossez et al. 2024; Zhang et al. 2023; Zhan et al. 2024) take LLMs as the backbone and extend them with speech interaction. In theory, it fully utilizes LLMs' strengths (e.g., intent understanding, instruction following, planning, and decision-making). However, the quality of speech generation often suffers due to challenges in cross-modal alignment. This work focuses on the *LLM-centric* SLMs, aiming to explore how core components affect cross-modal alignment and the quality of generated speech.

The **speech tokenizer** and **speech head** determine the atomic inputs and latent representations of speech signals, which are critical to cross-modal alignment. Existing **speech tokenizers** can be categorized into three types: *coupled* tokenizers (e.g., WavTokenizer (2025)) that jointly encode semantic and acoustic details; *semi-decoupled* tokenizers (e.g., SpeechTokenizer (2024)) that distill semantic content via HuBERT (2021) while retaining partial coupling; and *decoupled* tokenizers (e.g., FACodec (2024)) that separate speech into independent subspaces such as semantics, prosody, and timbre. Although all three achieve comparable performance in speech reconstruction tasks, their compatibility with SLMs remains underexplored. Under fair comparison within SLMs, we find that decoupled tokenizers are more favorable for cross-modal alignment and yield higher speech synthesis quality. We observe a mismatch in information density between speech and text representations in the **speech head**: one second of speech corresponds to hundreds of tokens,

* Equal contribution.
†Corresponding author.

far exceeding the amount of text information humans typically express per second. In LLMs, multi-token prediction (MTP) (Gloeckle et al. 2024; Cai et al. 2024; DeepSeek-AI 2025) is often used to accelerate inference or enrich hidden representations by enabling a single hidden state to predict multiple future tokens. To balance the information density between speech and text, we introduce MTP into SLMs, allowing each speech hidden vector to decode multiple speech tokens (for both coupled and decoupled tokenizers). We experimented with various compression ratios, ranging from one token per hidden state to twelve tokens per state. The MTP not only improved speech decoding speed by up to 12× but also significantly enhanced speech synthesis quality (the Word Error Rate decreased from 6.07 to 3.01).

In addition, we propose a speaker-aware speech generation paradigm that introduces speaker identity features into the speech-language context to guide and control the speaker's timbre in synthesized speech. To comprehensively evaluate SLMs' understanding and generalization of the general knowledge captured by LLMs, as well as their speaker consistency, we construct a role-playing knowledge QA task. Specifically, we use CosyVoice2 (2024b) to convert answers from a text-based knowledge QA dataset—TriviaQA (2017) into speech, assigning each answer to one of 15 Genshin characters. The resulting RoleTriviaQA dataset contains 138K/0.3K/2.4K samples for training/validation/test, respectively. We reserve five characters and the WebQuestions (2025) test set for out-of-domain evaluation of both speaker consistency and knowledge QA performance (i.e., evaluating generalization). Results show that the decoupled architecture enables cooperative improvement in knowledge QA tasks and speaker timbre, and significantly outperforms other baselines.

2 Speech-Language Model

A typical speech-language model (SLM) is usually an extension of a successful large language model (LLM) capable of understanding and generating both text and speech signals. It relies on a *speech-language tokenizer* to convert raw text or audio into discrete tokens, a *decoder-only Transformer* to learn cross-modal context and alignment, a *language head* to predict text, and a *speech head* to predict speech tokens and then synthesize continuous audio signals.

In the following, we introduce these core components and discuss how different design choices affect the SLM's capabilities, particularly in the speech-language tokenizer and speech head.

2.1 Core Components

Speech-Language Tokenizer is a key component of SLMs, as it defines the atomic representations of text and speech. High-quality representations are crucial for effective cross-modal alignment, efficient generation, and preservation of acoustic details, making this component the focus of extensive research (Zhan et al. 2024; Peng et al. 2024). It converts text or speech into discrete token sequences, such as $X_{\text{📄}} = \{t_1, t_2, \dots, t_{|\text{📄}|}\}$ for text and $X_{\text{🔊}} = \{s_1, s_2, \dots s_{|\text{🔊}|}\}$ for speech. These sequences are ultimately combined into interleaved text-speech sequences $X = \{\dots, X_{\text{🔊}}, X_{\text{📄}}, X_{\text{🔊}}, \dots\}$ to handle a variety of speech-language tasks, such as ASR (Fathullah et al. 2024), TTS (Du et al. 2024a,b), speech translation (Chen et al. 2024), and spoken question answering (Zhao et al. 2024).

Speech tokenizers can be categorized into coupled $X_{\text{🔊}} = \{s_1, s_2, \dots\}$ and decoupled $X_{\text{🔊}} = \{\bar{s}_1, \tilde{s}_1, \bar{s}_2, \tilde{s}_2, \dots\}$ [1, 2] designs for semantic $\bar{s}_*$ and acoustic details $\tilde{s}_*$. Intuitively, as illustrated in Figure 1, decoupled semantic tokens are easier to align with text tokens, while the SLM can leverage its overparameterization to handle the acoustic detail tokens specifically. In contrast, coupled tokenizers may compromise cross-modal semantic alignment.

However, current comparisons between speech tokenizers are primarily conducted on speech reconstruction tasks, which are independent of SLMs. As a result, it is difficult to assess their suitability for SLMs accurately. For example, under fair conditions, the coupled BigCodec (2024) achieves better reconstruction quality, yet when integrated into an SLM, the decoupled FACodec (2024) yields significantly better performance. This paper provides a detailed analysis of mainstream speech tokenizers within the context of SLM evaluation.

Decoder-only Transformer has achieved great success as a foundation model for text understanding and generation. After unifying the tokenization of text and speech, it can further learn joint speech-text context and perform cross-modal alignment. Its parameters are typically initialized from a well-pretrained LLM and then adapted to multimodal contexts. Given a sequence of input tokens X, it produces a hidden representation vector $\boldsymbol{h}_i \in \mathbb{R}^d$, $H = \{\boldsymbol{h}_1, \boldsymbol{h}_2, \dots, \boldsymbol{h}_{|X|}\}$ and for each token, $\boldsymbol{h}_i = \mathrm{Transformer}(X_{\leq i})$.

Prediction Heads include a *language head* and a *speech head*. The language head is consistent with that of an LLM and can be initialized from pretrained parameters. In contrast, the speech head has a different "vocabulary size" and is randomly initialized. **Language head** consists of a linear transformation followed by a softmax function,

$$P\left(t_{i+1} \mid X_{\leq i}\right) = \mathrm{softmax}_{t_{i+1}}(W_{\text{📄}} \cdot \boldsymbol{h}_i). \qquad (1)$$

A parameter matrix $W_{\text{📄}} \in \mathbb{R}^{|V_{\text{📄}}| \times d}$ maps the hidden vector $\boldsymbol{h}_i$ to the vocabulary space of size $|V_{\text{📄}}|$. During training, the model is optimized using the cross-entropy loss: $\mathcal{L}_{\text{📄}} = -\log P\left(t_{i+1} \mid X_{\leq i}\right)$, and during inference, the next token is predicted by either taking the $\arg\max P(t_{i+1} \mid X_{\leq i})$ or sampling. **Speech head** follows a similar formal definition,

[1]For simplicity, we assume a two-way decomposition, though in practice it can be readily extended to more components, e.g., $\hat{s}_i, \check{s}_i, \dots$.

[2]Decoupled sequences may also be arranged in a type-wise contiguous manner within a chunk (e.g., $X_{\text{🔊}} = \{\bar{s}_1, \bar{s}_2, \cdots, \tilde{s}_1, \tilde{s}_2, \dots\}$) rather than in an interleaved fashion. However, experiments in Appendix F.3 (See the extended version) show that interleaved ordering yields significantly better performance.

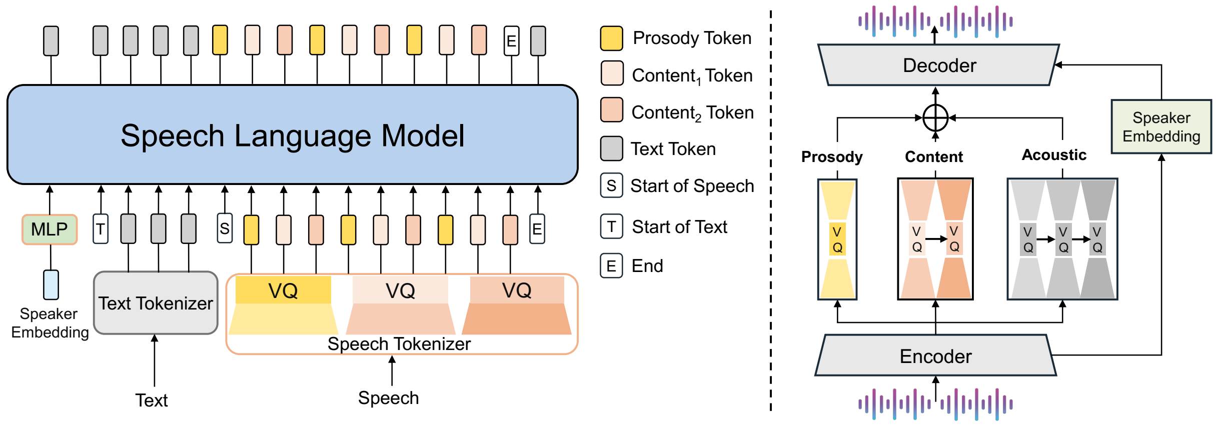

Figure 1: **Left**: Overview of a Speech Language Model (SLM) trained with a decoupled speech tokenizer (Section 2.1) and Speaker-Aware TTS (Section 2.3); **Right**: The architecture of a possible decoupled speech tokenizer, featuring speech quantization, reconstruction in a decoupled manner, and speaker-specified embedding extraction.

training objective, and inference process as the language head,

$$P(s_{i+1} \mid X_{\le i}) = \text{softmax}_{\ s_{i+1}}(W_{\text{🔊}} \cdot \boldsymbol{h}_i). \tag{2}$$

It uses a parameter matrix $W_{\text{🔊}} \in \mathbb{R}^{|V_{\text{🔊}}| \times d}$ to map the hidden vector $\boldsymbol{h}_i$ to the speech "vocabulary" of size $|V_{\text{🔊}}|$. When semantic and acoustic components are decoupled (i.e., s_i is split into $\bar{s}_i$ and $\tilde{s}_i$), the vocabulary and the linear projection are also split accordingly into independent parts:

$$\begin{cases} P(\bar{s}_{i+1} \mid X_{\le i}) = \text{softmax}_{\ \bar{s}_{i+1}}(\overline{W}_{\text{🔊}} \cdot \boldsymbol{h}_i) \\ P(\tilde{s}_{i+1} \mid X_{\le i}) = \text{softmax}_{\ \tilde{s}_{i+1}}(\widetilde{W}_{\text{🔊}} \cdot \boldsymbol{h}_i), \end{cases} \tag{3}$$

where $\overline{W}_{\text{🔊}} \in \mathbb{R}^{|\overline{V}_{\text{🔊}}| \times d}$ and $\widetilde{W}_{\text{🔊}} \in \mathbb{R}^{|\widetilde{V}_{\text{🔊}}| \times d}$.

Finally, the speech is reconstructed through the audio decoder from a pretrained codec: $\text{🔊} = \text{Decoder}(X_{\text{🔊}})$.

2.2 Multi Token Prediction

We observe that text representations are significantly more information-dense than speech. Typically, one second of speech is decoded into several hundred tokens, while it rarely conveys more than ∼20 text tokens' worth of information. This asymmetry in information density can make cross-modal alignment difficult to learn. On the other hand, speech tokens with excessively low information density can also hinder SLM inference efficiency, as generating one second of speech may take significantly longer than one second. Therefore, balancing the information density between speech and text representations becomes a critical challenge.

Multi-token prediction (MTP) (Gloeckle et al. 2024; Cai et al. 2024; DeepSeek-AI 2025) is commonly used in LLMs to accelerate inference or enrich hidden representations (e.g., by requiring $\boldsymbol{h}_i$ to decode not only t_{i+1} but also future tokens like $t_{i+2}, \ldots$). To compress the information density of speech representations, we introduce MTP into the SLM framework. As depicted in Figure 2, compared with Next Token Prediction (NTP), MTP enables each hidden vector $\boldsymbol{h}_i$ to decode multiple speech tokens, whether coupled or decoupled, thus alleviating the imbalance in information density between speech and text.

We group every g adjacent speech tokens into a multi-token group G_j, which is decoded simultaneously at step j, meaning all tokens in G_j are predicted from the same hidden vector $\boldsymbol{h}_j$. This significantly increases the information density encoded in $\boldsymbol{h}_j$, making it more compatible with text tokens. The definitions of the multi-token groups $G_{j+1}^{(coupled)}$/$G_{j+1}^{(decoupled)}$ [3] for coupled/decoupled tokenizers subject to $\left|G_{j+1}^{(*)}\right| = g$ are as follows:

$$\begin{aligned} G_{j+1}^{(coupled)} &= \{s_{j\times g+1}, s_{j\times g+2}, \ldots, s_{j\times g+g}\} \\ G_{j+1}^{(decoupled)} &= \{\bar{s}_{j\times \frac{g}{2}+1}, \tilde{s}_{j\times \frac{g}{2}+1}, \ldots, \bar{s}_{j\times \frac{g}{2}+\frac{g}{2}}, \tilde{s}_{j\times \frac{g}{2}+\frac{g}{2}}\}. \end{aligned}$$

The parameter matrix of the speech head is extended from a 2D tensor to a 3D tensor to enable simultaneous prediction of multiple tokens. The extended dimension has size g, where each slice corresponds to a linear projection matrix (as in Equation (2) for coupled or Equation (3) for decoupled tokens) for one token in the group G,

$$\begin{aligned} \boldsymbol{W}_{\text{🔊}}^{(coupled)} &= [W_{\text{🔊}}^1, W_{\text{🔊}}^2, \ldots, W_{\text{🔊}}^g], \\ \boldsymbol{W}_{\text{🔊}}^{(decoupled)} &= [\overline{W}_{\text{🔊}}^1, \widetilde{W}_{\text{🔊}}^1, \ldots, \overline{W}_{\text{🔊}}^{g/2}, \widetilde{W}_{\text{🔊}}^{g/2}]. \end{aligned}$$

Similarly, the softmax normalization is applied after the linear transformations to yield the probability distributions for all tokens in the group,

$$P(G_{j+1}^{(*)} \mid X_{\le j\times g}) = \text{softmax}_{G_{j+1}^{(*)}}(\boldsymbol{W}_{\text{🔊}}^{(*)} \cdot \boldsymbol{h}_j).$$

[3] Where the group $G_{j+1}^{(decoupled)}$ consists of half semantic tokens $\bar{s}_*$ and half acoustic detail tokens $\tilde{s}_*$.

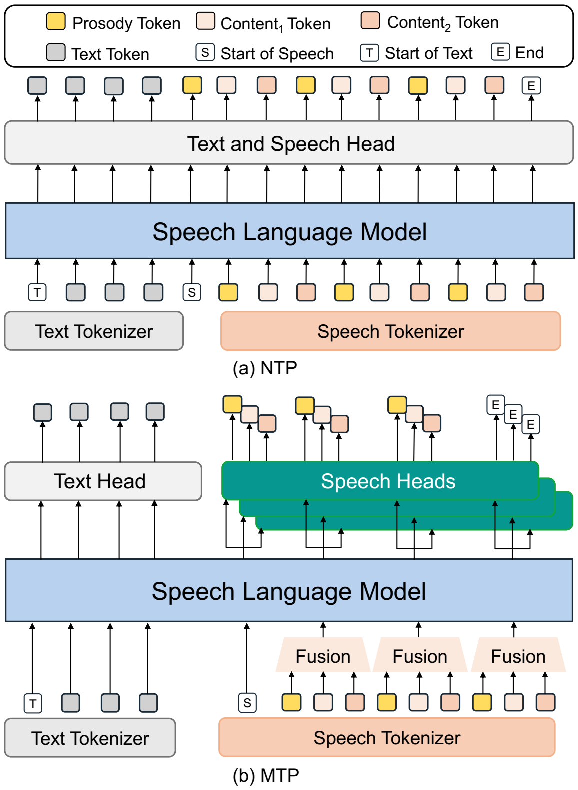

Figure 2: Illustration of our NTP and MTP architecture. (a) NTP: single vocabulary and single prediction head; (b) MTP: multiple vocabularies and multiple prediction heads, generating multiple tokens in parallel.

The loss $\mathcal{L}_{🔊}^{(coupled)}/\mathcal{L}_{🔊}^{(decoupled)}$ is also extended to the average over all tokens within the group: $\mathcal{L}_{🔊}^{(*)} = \frac{1}{g}\sum_{s_k \in G_{j+1}^{(*)}} -\log P(s_k \mid X_{\leq j \times g})$.

Finally, the Transformer input must be adapted for MTP. For each group, we first obtain the embeddings of all tokens (i.e., $\boldsymbol{s}_j = \mathrm{E}(s_j)$ in NTP), concatenate them, and feed the result into a fusion network like a linear or MLP to produce a downsampled (by a factor of $^1/_g$) input vector:

$$\boldsymbol{s}_j = \underbrace{\mathrm{Fusion}}_{e.g.,\mathrm{MLP}}\Big(\bigoplus_{s_k \in G_j^{(*)}} \mathrm{E}(s_k)\Big), \qquad (4)$$

where the $\bigoplus$ is the concatenation function.

2.3 Speaker-Aware Speech Generation

In speech generation (especially within spoken dialogue systems), explicitly controlling paralinguistic features such as speaker timbre, intonation, and emotional tone enables the generation of consistent and controllable speech responses. Therefore, we introduce speaker-specific representations into the SLM context and evaluate both speech quality and character consistency in a role-playing speech QA task.

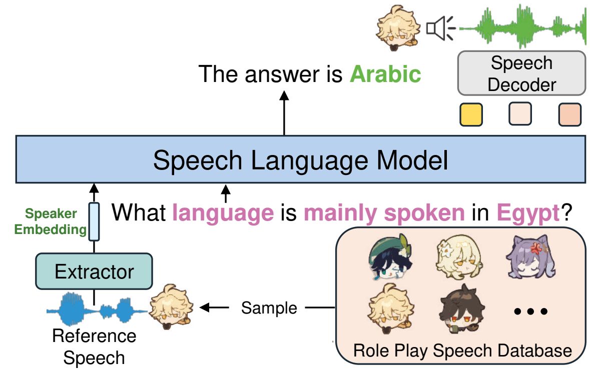

Figure 3: Illustration of Role-Playing Knowledge QA.

Inspired by Du et al. (2024a) and Li et al. (2024), we employ a pretrained timbre extractor [4] to obtain the timbre representation $X_{👤}$ of the speaker or role. This representation can either be a discretized token $X_{👤} = \{u_1, u_2, \ldots, u_{|👤|}\}$, or a continuous vector sequence $X_{👤} = \{\boldsymbol{u}_1, \boldsymbol{u}_2, \ldots, \boldsymbol{u}_{|👤|}\}$, where $|👤|$ is the sequence length. The speech-language context of the SLM is then extended as follows: $X = \{X_{👤}, \ldots, X_{🔊}, X_{📄}, X_{🔊}, \ldots\}$.

We expect the generated speech to be guided and controlled by the speaker-specific features.

3 Role-Playing Knowledge QA Task

Current SLMs excel at daily conversations (Défossez et al. 2024), but their performance on knowledge-intensive tasks (e.g., QA) and voice characteristics (e.g., role-playing) lacks systematic evaluation. To address this problem, we introduce the **Role-Playing Knowledge QA** task, assessing SLM knowledge retention from its backbone LLM and similarity to reference speech in voice characteristics.

However, a critical gap exists: **no open-source dataset** assesses both correctness and timbre similarity in speech knowledge QA. Other models featuring speech QA (Défossez et al. 2024) were trained and evaluated on a large-scale dataset, which has not been released yet.

To bridge this gap and stimulate SLMs' intrinsic knowledge with substantial training (Zhang et al. 2023), we propose **RoleTriviaQA**, an open-source dataset for benchmarking SLMs on role embodiment and knowledgeable spoken responses. Similar with (Zhang et al. 2023), we adopt the data format $X^{(role\text{-}qa)} = \{X_{👤}, X_{📄}^{(q)}, X_{📄}^{(a)}, X_{🔊}^{(a)}\}$ (see Figure 3), which enables essential text-speech knowledge alignment for high-quality role-playing.

3.1 Data Construction

We created the RoleTriviaQA dataset by selecting 15 distinct voices from *Genshin Impact* roles (10 seen: 5 Male/5 Female for training and evaluation; 5 unseen: 2 Male/3 Female

[4] We adopt the timbre extractor from the decoupled tokenizer FACodec (Ju et al. 2024) in this work. Implementation details are provided in Appendix B. It can be smoothly replaced with other extractors if needed.

Model	Codebook Size	Nq	Token Rate (TPS)	Success Rate (SR) ↑	UTMOS ↑	WER ↓	SIM ↑	
							Seen	Unseen
Coupled Tokenizers trained with SLM								
Encodec (2023)	1024	4	300	0.38	2.86	17.00	0.18	0.10
Encodec (2023)	1024	8	600	0.24	3.50	8.14	0.18	0.10
WavTokenizer (2025)	4096	1	75	0.38	2.80	89.60	0.06	0.04
WavTokenizer-v2 (2025)	4096	1	75	0.65	3.16	49.24	0.11	0.08
StableCodec (2025)	15625	2	50	0.98	3.90	13.04	0.17	0.14
BigCodec (2024)	8192	1	80	0.70	3.93	13.63	0.18	0.10
Semi-Decoupled Tokenizers trained with SLM								
SpeechTokenizer (2024)	1024	5	250	0.99	3.86	13.13	0.16	0.13
SpeechTokenizer (2024)	1024	8	400	0.90	**4.05**	9.95	0.14	0.14
Decoupled Tokenizers trained with SLM								
FACodec-NTP-1H	1024	3	240	0.87	3.93	6.07	**0.50**	**0.49**
FACodec-MTP-3H	1024	3	80	0.82	3.83	5.84	0.47	0.47
FACodec-MTP-6H	1024	3	40	0.92	3.75	4.36	0.47	0.47
FACodec-MTP-12H	1024	3	20	**1.00**	3.67	**3.01**	0.47	0.47

Table 1: Comparison of coupled, semi-decoupled, and decoupled tokenizers trained with SLMs (H: speech heads for FACodec-based SLMs). Evaluated on LibriTTS Test-Clean, with SIM tested on both seen (Train) and unseen (Test-Clean) speakers. Best and second-best results are highlighted in **bold** and underline, respectively. Detailed information about these baselines can be found in Appendix A.1.

Model	H	TPS	Spk-Aware	UTMOS↑	WER↓	SIM↑
NTP	1	240	✗	3.93	6.07	0.49
NTP	1	240	✓	**4.05**	5.51	0.57
MTP	3	80	✗	3.83	5.84	0.47
MTP	3	80	✓	3.90	7.04	0.56
MTP	6	40	✗	3.75	4.36	0.47
MTP	6	40	✓	3.94	4.07	0.59
MTP	12	20	✗	3.67	3.01	0.47
MTP	12	20	✓	3.93	**2.94**	**0.60**

Table 2: Impact of Speaker-Aware training for decoupled tokenizer with NTP and MTP architectures, evaluated on LibriTTS test-clean dataset.

for evaluation only). Then we synthesize speech for TriviaQA (Joshi et al. 2017) QA-pairs using CosyVoice 2 (Du et al. 2024b) after filtering high-quality reference speeches. This pipeline yields 138,384 training samples and a test set with 1,500 in-domain (ID, from TriviaQA) and 926 out-of-domain (OOD, from OpenAudioBench (Li et al. 2025)) samples. Further details of RoleTriviaQA are in Appendix H.

3.2 Training Strategy

To build an SLM support role-playing knowledge QA task, we adopt a two-stage training method:

Stage 1: Cross-Modal Alignment Pretraining To improve cross-modal alignment, we conduct a large-scale dataset **Emilia-3.5**. First, we filter Emilia (2024) (a large-scale, real-world speech collection with diverse emotions and styles) for samples scoring at least 3.5 on DNSMOS (2022). Next, we merge the standard LibriTTS (2019) training set. The resulting corpus comprises 2,495 hours of high-quality speech. Moreover, to enable bidirectional alignment of text and speech modalities, we conduct pre-training on both TTS and ASR tasks. Specifically, we adopt $X^{(tts)} = \{X_{👤}, X_{📄}, X_{🔊}\}$ as the TTS data format and $X^{(asr)} = \{X_{🔊}, X_{📄}\}$ for ASR.

The pretraining loss $\mathcal{L}^{(pt)}$ is formulated as: $\mathcal{L}^{(pt)} = \sum_{y_i \in X^{(tts/asr)}} -\log P(y_i \mid X^{(tts/asr)}_{<i})$, where the predicted token y_i could be a single token or a multi-token group (i.e., G_i).

Stage 2: Knowledge-Role Joint Fine-Tuning After the pretraining stage, we perform supervised fine-tuning (SFT) with RoleTriviaQA to enable the SLM to capture the voice features of each role along with the knowledge-intensive data.

The SFT loss $\mathcal{L}^{(role\text{-}qa)}$ is defined as follows: $\mathcal{L}^{(role\text{-}qa)} = \sum_{y_i \in X^{(a)}_{📄 \cup 🔊}} -\log P(y_i \mid X^{(role\text{-}qa)}_{<i})$, where the loss is calculated across both the text answer and the speech answer.

4 Experiment

Our experiments address two critical questions:

1. How do **different speech tokenizer** and **training methods** (NTP/MTP/Speaker-Aware) affect SLMs' performance? (Sections 4.1-4.3);
2. How can SLMs effectively leverage their intrinsic knowledge while maintaining high-quality speech synthesis? (Section 4.4).

Setups Details of the datasets, baselines, evaluation, and hyperparameter setups are placed in Appendix C.

4.1 Effect of Speech Tokenizers

As shown in Table 1, the degree of speech representation decoupling shows a significant correlation with the joint SLM training effect, where the semi-decoupled and fully decoupled architectures exhibit better overall performance.

Models	Model Type	Tokenizer Type	In-Domain (RoleTriviaQA)					Out-of-Domain (Web Questions)				
			EM(%)↑	F1(%)↑	SIM↑ Seen	SIM↑ Unseen	SIM↑ All	EM↑	F1↑	SIM↑ Seen	SIM↑ Unseen	SIM↑ All
Qwen2.5-0.5B-Instruct	LLM	-	7.1	17.3	-	-	-	1.1	14.0	-	-	-
WavTokenizer-v2	SLM	Coupled	5.1	12.5	0.22	<u>0.18</u>	0.20	<u>1.1</u>	3.9	0.21	<u>0.18</u>	0.20
BigCodec	SLM	Coupled	2.9	9.4	<u>0.23</u>	0.17	<u>0.21</u>	1.0	<u>4.2</u>	<u>0.23</u>	0.17	<u>0.21</u>
StableCodec	SLM	Coupled	<u>7.5</u>	<u>13.8</u>	0.18	0.14	0.17	0.6	3.6	0.17	0.14	0.16
SpeechTokenizer	SLM	Semi-Decoupled	3.4	9.1	0.19	0.14	0.17	0.3	2.6	0.20	0.14	0.18
FACodec	SLM	Decoupled	**12.0**	**23.8**	**0.63**	**0.58**	**0.61**	**9.0**	**17.9**	**0.63**	**0.58**	**0.61**

Table 3: Comparisons of coupled, semi-decoupled, and decoupled tokenizers trained with SLMs on the RoleTriviaQA test set.

First, coupled architectures exhibit severe convergence challenges during SLM training (see Appendix C), resulting in generally low success rates of speech synthesis ($\leq$70%). The notable exception is StableCodec, whose large codebook and low token rate establish dual compensation mechanisms that mitigate inherent limitations of coupled architectures. Second, the semi-decoupled and decoupled architectures show clear advantages. The semi-decoupled architecture achieves the best speech quality, while the decoupled architecture works optimally in WER and SIM. Third, the decoupled architectures' SIM between the training and test sets is consistently larger and significantly better than the other baselines, suggesting superior generalization capabilities, especially for speaker timbre.

Overall, decoupled tokenizer and SLM joint modeling is better and has better controllability and scalability.

4.2 Next Token vs. Multi-Token Prediction

As illustrated in Table 1, to investigate the effectiveness of MTP under decoupled architecture, we also conduct experiments comparing models trained with different compression ratios.

First, the MTP-3H model is competitive with the NTP one, even with 3$\times$ compression. Moreover, it has also surpassed all of the semi-decoupled and coupled baselines in WER and SIM. Second, when increasing compression from 3$\times$ (MTP-3H) to 12$\times$ (MTP-12H), WER improves by 48% (5.84 to 3.01), with stable SIM and UTMOS experiencing a slight decline. Notably, under 12$\times$ compression, our model achieves speech synthesis with a 100% success rate, and the WER (3.01) is comparable to the ground truth. These results reveal that speech length compression based on the decoupled tokenizer effectively mitigates the frequency mismatch between text ($\sim$20Hz) and speech (240Hz) modalities, enhancing cross-modal alignment performance.

4.3 Speaker-Aware Speech Generation

Table 2 presents that Speaker-Aware training exhibits consistent enhancement effects across NTP and MTP architectures with a decoupled tokenizer.

First, for NTP, the incorporation of speaker embedding leads to consistent improvements across all metrics. Second, for MTP, speaker-embedding provides a stable gain in most cases. In addition, as compression increases from 3H to 12H, Speaker-Aware models also show continuous WER reduction (7.04 to 2.94), In particular, the advantages of Speaker-Aware training become more evident at higher compression. For instance, the MTP-12H trained with speaker embedding significantly outperforms its non-aware model.

This indicates that incorporating speaker embeddings enables the model better to capture the speaker's paralinguistic features and linguistic accuracy.

4.4 Role-Playing Knowledge Speech QA

Settings We evaluate Role-Playing Knowledge Speech QA within two aspects: 1) Benchmark of the decoupled SLM against top-3 coupled/semi-decoupled SLMs (From Table 1) and the LLM backbone (Qwen2.5-0.5B-Instruct); 2) Compare the decoupled SLM against diverse SLM baselines.

Results on Different Tokenizers Table 3 shows that SLM trained with a decoupled tokenizer achieves SOTA performance across all metrics. For ID evaluation, it outperforms all baselines with 12.0 EM and 23.8 F1, while its 0.61 SIM score is 2.9$\times$ higher than the best coupled baseline (BigCodec). Though all models degrade on OOD queries, the decoupled model shows greater robustness, confirming that decoupling resolves target conflicts in semi-decoupled/coupled systems by jointly optimizing semantic and acoustic objectives.

Compared to the LLM baseline, the decoupled SLM significantly outperforms in both ID/OOD settings (OOD EM: 9.0 vs 1.1; F1: 17.9 vs 14.0), effectively preserving LLM knowledge while adapting to unseen queries. It also generalizes well in speaker timbre, achieving high SIM scores for both seen/unseen speakers (0.63 vs 0.58) with minimal gaps.

Overall, the decoupled architecture thus enables cooperative improvement in knowledge QA and speaker timbre, surpassing all baselines.

Results on Different SLMs Comparisons with other SLM baselines also confirm that decoupled tokenizers help not only improve the speech quality but also preserve LLM knowledge. Detailed results are provided in Appendix E.

4.5 Ablation Study

MTP Architecture Selection We analyze 6 different fusion strategies and speech head architectures to select the optimal MTP configuration (**MLP-Linear** combination), as elaborated in Appendix F.1.

Models	Heads	TPS	Cosine Similarity			Euclidean Distance			Riemannian Distance↓
			TTSim	SSSim	STSim↑	TTDist	SSDist	STDist↓	
NTP	1	240	0.896	0.999	0.480	0.438	0.030	1.014	365.26
MTP	3	80	0.986	0.999	0.967	0.147	0.052	0.244	288.39
MTP	6	40	0.991	0.998	<u>0.981</u>	0.116	0.055	<u>0.186</u>	<u>268.19</u>
MTP	12	20	0.996	0.998	**0.986**	0.078	0.063	**0.158**	**237.55**

Table 4: Cross-modal alignment metrics (calculated on the hidden states of the last layer) across different speech token compression rates. **Metrics**: TTSim/SSSim (intra-modal cosine similarity: text vs. speech); STSim (cross-modal cosine similarity); TTDist/SSDist (intra-modal Euclidean distance: text vs. speech); STDist (cross-modal Euclidean distance).

NTP Layer and Token Organization Selection We explore the best number of decoupled tokenizer codebook layers and speech token organization patterns for SLM training. Results show that **3 codebook layers** with **Frame-Wise Interleaving organization pattern** perform the best. Details are provided in Appendix F.2 and Appendix F.3.

RoleTriviaQA Data Format Selection We conducted experiments on 3 different data formats for training with RoleTriviaQA and found that the current data format we adopted (see Section 3) performs best across all baselines compared with others. This experiment is detailed in Appendix F.4.

4.6 Effect of MTP on Cross-modal Alignment

Table 1 and Table 2 show that higher speech token compression in the MTP architecture consistently improves performance, especially in WER. To investigate the hypothesis that higher compression improves modal alignment, we conduct detailed quantitative and qualitative analyses.

Quantitative Analysis As shown in Table 4, token compression improves cross-modal alignment and maintains intra-modal integrity, with MTP models outperforming NTP baselines. First, when compression increases from $3\times$ to $12\times$, intra-modal similarities (TTSim, SSSim) remain high in all settings, while cross-modal similarity (STSim) increases steadily from 0.967 to 0.986. Second, the cross-model distance (STDist) decreases significantly from 0.244 to 0.158 , while intra-modal distances (TTDist, SSDist) stay stable. These trends align with WER improvements, suggesting that stronger compression encourages more isomorphic and effective speech-text representations. Additionally, compared to NTP, all MTP settings exhibit better cross-modal alignment. This demonstrates that MTP not only maintains intra-modal coherence but also enhances cross-modal correspondence under token compression.

We also use Riemannian distance (2009) to quantify speech-text alignment (See Appendix D for details). Table 4 shows that with higher compression, the Riemannian distance decreases (288.39 to 237.55), indicating that aggressive compression not only reduces embedding distance but also better preserves their intrinsic geometric structure.

Qualitative Analysis We include complementary qualitative results in Appendix G.1.

4.7 Effect of Tokenizer-MTP Combination

To examine the effectiveness of different tokenizer coupling strategies in MTP SLMs, we perform comprehensive experiments in Appendix G.2.

5 Related Work

Speech Tokenization has advanced through RVQ-based architectures and quantization innovations: SoundStream (2022) introduced the RVQ-VAE framework, enhanced by Encodec (2023, LSTM temporal modeling) and SpeechTokenizer (2024, HuBERT semantic distillation). Recent ultra-low-bitrate methods include Improved RVQGAN (DAC, 2023) with quantizer dropout, WavTokenizer (2025, 0.9kbps via single-layer RVQ), BigCodec (2024, 1.04kbps via Convolution-LSTM projection), and StableCodec (2025, 0.7kbps using Transformer with FSQ (2024)).

Speech Language Models LLMs are increasingly adopted for unified speech processing through two primary paradigms. The first integrates speech encoders with LLMs for comprehension tasks like ASR (Ma et al. 2024) and spoken language understanding (Chu et al. 2023), achieving strong analytical capabilities but lacking generative functionality. The second targets speech generation via dual strategies: codec-based models (e.g., SpeechGPT, 2023) discretize speech into compressed tokens, while diffusion-enhanced approaches (e.g., CosyVoice, 2024a; 2024b) synthesize waveforms through LLM-diffusion.

Multi Token Prediction has emerged as a critical technique in LLMs to accelerate inference. Detailed related works are placed in Appendix I.

6 Conclusion

This work investigates how LLM-centric SLMs can better align and generalize speech-text understanding and generation by revisiting core design choices. We find that fully decoupled tokenizers and multi-token prediction significantly improve alignment efficiency and speech quality. Additionally, incorporating speaker-aware modeling and role-based QA evaluation reveals improved generalization and character consistency. Our findings provide practical insights for building more capable and scalable SLMs.

Acknowledgements

The authors wish to thank the anonymous reviewers for their helpful comments. This work was partially funded by National Natural Science Foundation of China (No.62476061, 62376061,62576106, 62506079), Shanghai Rising-Star Program (23QA1400200), and Natural Science Foundation of Shanghai (23ZR1403500). The computations in this research were performed using the CFFF platform of Fudan University.

References

Anastassiou, P.; Chen, J.; Chen, J.; et al. 2024. Seed-TTS: A Family of High-Quality Versatile Speech Generation Models. arXiv:2406.02430.

Bonnabel, S.; and Sepulchre, R. 2009. Riemannian Metric and Geometric Mean for Positive Semidefinite Matrices of Fixed Rank. *SIAM Journal on Matrix Analysis and Applications*, 31: 1055–1070.

Cai, T.; Li, Y.; Geng, Z.; Peng, H.; Lee, J. D.; Chen, D.; and Dao, T. 2024. Medusa: Simple LLM Inference Acceleration Framework with Multiple Decoding Heads. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net.

Chen, X.; Zhang, S.; Bai, Q.; Chen, K.; and Nakamura, S. 2024. LLaST: Improved End-to-end Speech Translation System Leveraged by Large Language Models. arXiv:2407.15415.

Chu, Y.; Xu, J.; Zhou, X.; Yang, Q.; Zhang, S.; Yan, Z.; Zhou, C.; and Zhou, J. 2023. Qwen-Audio: Advancing Universal Audio Understanding via Unified Large-Scale Audio-Language Models. arXiv:2311.07919.

DeepSeek-AI. 2025. DeepSeek-V3 Technical Report. arXiv:2412.19437.

Défossez, A.; Copet, J.; Synnaeve, G.; and Adi, Y. 2023. High Fidelity Neural Audio Compression. *Trans. Mach. Learn. Res.*, 2023.

Du, Z.; Chen, Q.; Zhang, S.; Hu, K.; Lu, H.; Yang, Y.; Hu, H.; Zheng, S.; Gu, Y.; Ma, Z.; Gao, Z.; and Yan, Z. 2024a. CosyVoice: A Scalable Multilingual Zero-shot Text-to-speech Synthesizer based on Supervised Semantic Tokens. arXiv:2407.05407.

Du, Z.; Wang, Y.; Chen, Q.; Shi, X.; Lv, X.; Zhao, T.; Gao, Z.; Yang, Y.; Gao, C.; Wang, H.; Yu, F.; Liu, H.; Sheng, Z.; Gu, Y.; Deng, C.; Wang, W.; Zhang, S.; Yan, Z.; and Zhou, J. 2024b. CosyVoice 2: Scalable Streaming Speech Synthesis with Large Language Models. arXiv:2412.10117.

Défossez, A.; Mazaré, L.; Orsini, M.; Royer, A.; Pérez, P.; Jégou, H.; Grave, E.; and Zeghidour, N. 2024. Moshi: A Speech-Text Foundation Model for Real-Time Dialogue. arXiv:2410.00037.

Fathullah, Y.; Wu, C.; Lakomkin, E.; Jia, J.; Shangguan, Y.; Li, K.; Guo, J.; Xiong, W.; Mahadeokar, J.; Kalinli, O.; Fuegen, C.; and Seltzer, M. 2024. Prompting Large Language Models with Speech Recognition Abilities. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2024, Seoul, Republic of Korea, April 14-19, 2024*, 13351–13355. IEEE.

Gloeckle, F.; Idrissi, B. Y.; Rozière, B.; Lopez-Paz, D.; and Synnaeve, G. 2024. Better & Faster Large Language Models via Multi-token Prediction. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net.

Grattafiori, A.; Dubey, A.; Jauhri, A.; et al. 2024. The Llama 3 Herd of Models. arXiv:2407.21783.

He, H.; Shang, Z.; Wang, C.; Li, X.; Gu, Y.; Hua, H.; Liu, L.; Yang, C.; Li, J.; Shi, P.; Wang, Y.; Chen, K.; Zhang, P.; and Wu, Z. 2024. Emilia: An Extensive, Multilingual, and Diverse Speech Dataset For Large-Scale Speech Generation. In *IEEE Spoken Language Technology Workshop, SLT 2024, Macao, December 2-5, 2024*, 885–890. IEEE.

Hsu, W.; Bolte, B.; Tsai, Y. H.; Lakhotia, K.; Salakhutdinov, R.; and Mohamed, A. 2021. HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units. *IEEE ACM Trans. Audio Speech Lang. Process.*, 29: 3451–3460.

Ji, S.; Jiang, Z.; Wang, W.; Chen, Y.; Fang, M.; Zuo, J.; Yang, Q.; Cheng, X.; Wang, Z.; Li, R.; Zhang, Z.; Yang, X.; Huang, R.; Jiang, Y.; Chen, Q.; Zheng, S.; and Zhao, Z. 2025. WavTokenizer: an Efficient Acoustic Discrete Codec Tokenizer for Audio Language Modeling. In *The Thirteenth International Conference on Learning Representations*.

Joshi, M.; Choi, E.; Weld, D. S.; and Zettlemoyer, L. 2017. TriviaQA: A Large Scale Distantly Supervised Challenge Dataset for Reading Comprehension. In Barzilay, R.; and Kan, M., eds., *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers*, 1601–1611. Association for Computational Linguistics.

Ju, Z.; Wang, Y.; Shen, K.; Tan, X.; Xin, D.; Yang, D.; Liu, E.; Leng, Y.; Song, K.; Tang, S.; Wu, Z.; Qin, T.; Li, X.; Ye, W.; Zhang, S.; Bian, J.; He, L.; Li, J.; and Zhao, S. 2024. NaturalSpeech 3: Zero-Shot Speech Synthesis with Factorized Codec and Diffusion Models. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net.

Kim, M. J.; Pertsch, K.; Karamcheti, S.; Xiao, T.; Balakrishna, A.; Nair, S.; Rafailov, R.; Foster, E. P.; Sanketi, P. R.; Vuong, Q.; Kollar, T.; Burchfiel, B.; Tedrake, R.; Sadigh, D.; Levine, S.; Liang, P.; and Finn, C. 2024. OpenVLA: An Open-Source Vision-Language-Action Model. In Agrawal, P.; Kroemer, O.; and Burgard, W., eds., *Conference on Robot Learning, 6-9 November 2024, Munich, Germany*, volume 270 of *Proceedings of Machine Learning Research*, 2679–2713. PMLR.

Kumar, R.; Seetharaman, P.; Luebs, A.; Kumar, I.; and Kumar, K. 2023. High-Fidelity Audio Compression with Improved RVQGAN. In Oh, A.; Naumann, T.; Globerson, A.; Saenko, K.; Hardt, M.; and Levine, S., eds., *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Li, T.; Liu, J.; Zhang, T.; Fang, Y.; Pan, D.; Wang, M.; Liang, Z.; Li, Z.; Lin, M.; Dong, G.; Xu, J.; Sun, H.; Zhou, Z.; and

Chen, W. 2025. Baichuan-Audio: A Unified Framework for End-to-End Speech Interaction. arXiv:2502.17239.

Li, Y. A.; Jiang, X.; Darefsky, J.; Zhu, G.; and Mesgarani, N. 2024. Style-Talker: Finetuning Audio Language Model and Style-Based Text-to-Speech Model for Fast Spoken Dialogue Generation. arXiv:2408.11849.

Liu, H.; Li, C.; Wu, Q.; and Lee, Y. J. 2023. Visual Instruction Tuning. In Oh, A.; Naumann, T.; Globerson, A.; Saenko, K.; Hardt, M.; and Levine, S., eds., *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Ma, Z.; Yang, G.; Yang, Y.; Gao, Z.; Wang, J.; Du, Z.; Yu, F.; Chen, Q.; Zheng, S.; Zhang, S.; and Chen, X. 2024. An Embarrassingly Simple Approach for LLM with Strong ASR Capacity. arXiv:2402.08846.

Mentzer, F.; Minnen, D.; Agustsson, E.; and Tschannen, M. 2024. Finite Scalar Quantization: VQ-VAE Made Simple. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Parker, J. D.; Smirnov, A.; Pons, J.; Carr, C.; Zukowski, Z.; Evans, Z.; and Liu, X. 2025. Scaling Transformers for Low-Bitrate High-Quality Speech Coding. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net.

Peng, P.; Huang, P.; Li, S.; Mohamed, A.; and Harwath, D. 2024. VoiceCraft: Zero-Shot Speech Editing and Text-to-Speech in the Wild. In Ku, L.; Martins, A.; and Srikumar, V., eds., *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, 12442–12462. Association for Computational Linguistics.

Qwen. 2025. Qwen2.5 Technical Report. arXiv:2412.15115.

Reddy, C. K. A.; Gopal, V.; and Cutler, R. 2022. Dnsmos P.835: A Non-Intrusive Perceptual Objective Speech Quality Metric to Evaluate Noise Suppressors. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2022, Virtual and Singapore, 23-27 May 2022*, 886–890. IEEE.

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

Xu, J.; Guo, Z.; He, J.; Hu, H.; He, T.; Bai, S.; Chen, K.; Wang, J.; Fan, Y.; Dang, K.; Zhang, B.; Wang, X.; Chu, Y.; and Lin, J. 2025. Qwen2.5-Omni Technical Report. arXiv:2503.20215.

Zeghidour, N.; Luebs, A.; Omran, A.; Skoglund, J.; and Tagliasacchi, M. 2022. SoundStream: An End-to-End Neural Audio Codec. *IEEE ACM Trans. Audio Speech Lang. Process.*, 30: 495–507.

Zen, H.; Dang, V.; Clark, R.; Zhang, Y.; Weiss, R. J.; Jia, Y.; Chen, Z.; and Wu, Y. 2019. LibriTTS: A Corpus Derived from LibriSpeech for Text-to-Speech. In Kubin, G.; and Kacic, Z., eds., *20th Annual Conference of the International Speech Communication Association, Interspeech 2019, Graz, Austria, September 15-19, 2019*, 1526–1530. ISCA.

Zhan, J.; Dai, J.; Ye, J.; Zhou, Y.; Zhang, D.; Liu, Z.; Zhang, X.; Yuan, R.; Zhang, G.; Li, L.; Yan, H.; Fu, J.; Gui, T.; Sun, T.; Jiang, Y.; and Qiu, X. 2024. AnyGPT: Unified Multimodal LLM with Discrete Sequence Modeling. In Ku, L.; Martins, A.; and Srikumar, V., eds., *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, 9637–9662. Association for Computational Linguistics.

Zhang, D.; Li, S.; Zhang, X.; Zhan, J.; Wang, P.; Zhou, Y.; and Qiu, X. 2023. SpeechGPT: Empowering Large Language Models with Intrinsic Cross-Modal Conversational Abilities. In Bouamor, H.; Pino, J.; and Bali, K., eds., *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*, 15757–15773. Association for Computational Linguistics.

Zhang, X.; Zhang, D.; Li, S.; Zhou, Y.; and Qiu, X. 2024. SpeechTokenizer: Unified Speech Tokenizer for Speech Language Models. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Zhao, Z.; Jiang, Y.; Liu, H.; Wang, Y.; and Wang, Y. 2024. LibriSQA: A Novel Dataset and Framework for Spoken Question Answering with Large Language Models. arXiv:2308.10390.