The Fortieth AAAI Conference on Artificial Intelligence (AAAI-26)

DualSpeechLM: Towards Unified Speech Understanding and Generation via Dual Speech Token Modeling with Large Language Models

Yuanyuan Wang[1], **Dongchao Yang**[1], **Yiwen Shao**[2], **Hangting Chen**[2], **Jiankun Zhao**[1],
Zhiyong Wu[1,3,*], **Helen Meng**[1], **Xixin Wu**[1,*]
[1]The Chinese University of Hong Kong
[2]Tencent AI Lab
[3]Tsinghua University

Abstract

Extending pre-trained text Large Language Models (LLMs)'s speech understanding or generation abilities by introducing various effective speech tokens has attracted great attention in the speech research community. However, building a unified speech understanding and generation model still faces the following challenges: (1) Due to the huge modality gap between speech and text tokens, extending text LLMs to unified speech LLMs relies on large-scale paired data for fine-tuning, and (2) Generation and understanding tasks prefer information at different levels, e.g., generation benefits from detailed acoustic features, while understanding favors high-level semantics. This divergence leads to difficult performance optimization in one unified model. To solve these challenges, in this paper, we present two key insights in speech tokenization and speech language modeling. Specifically, we first propose an Understanding-driven Speech Tokenizer (USTokenizer), which extracts high-level semantic information essential for accomplishing understanding tasks using text LLMs. In this way, USToken enjoys better modality commonality with text, which reduces the difficulty of modality alignment in adapting text LLMs to speech LLMs. Secondly, we present DualSpeechLM, a dual-token modeling framework that concurrently models USToken as input and acoustic token as output within a unified, end-to-end framework, seamlessly integrating speech understanding and generation capabilities. Furthermore, we propose a novel semantic supervision loss and a Chain-of-Condition (CoC) strategy to stabilize model training and enhance speech generation performance. Experimental results demonstrate that our proposed approach effectively fosters a complementary relationship between understanding and generation tasks, highlighting the promising strategy of mutually enhancing both tasks in one unified model.

Code — https://github.com/lavendery/UUG

Introduction

Recent advancements in autoregressive large language models (LLMs) have demonstrated excellent performance in the natural language processing community (Achiam et al. 2023; Touvron et al. 2023; Team et al. 2023). Leveraging the powerful foundations of text LLMs, recent advancements have led to emergence of speech LLMs that possess speech understanding and generation capabilities. In addition to fine-tuning textual LLMs to separately perform speech understanding tasks (Gong et al. 2023; Tang et al. 2023; Chu et al. 2023; Wang et al. 2024a, 2025) and generation tasks (Wang et al. 2023; Anastassiou et al. 2024; Kim et al. 2024; Wang et al. 2024c; Yang et al. 2024b, 2025a; Jia et al. 2025), developing unified models that excel at both capabilities has been explored in recent years (Zhang et al. 2023a; Xie et al. 2024; Fu et al. 2025; Nguyen et al. 2025; Défossez et al. 2024; Xu et al. 2025). However, several limitations still remain.

First, adapting pre-trained text LLMs to unified speech LLMs still relies heavily on large-scale paired speech-text data (Zhang et al. 2023a; Défossez et al. 2024; Xie et al. 2024; Xu et al. 2025). For example, SpeechGPT (Zhang et al. 2023a) and SpiritLM (Nguyen et al. 2025) require approximately 70K and 570K hours of paired data, respectively. This dependence stems from the substantial modality gap between speech and text, which hinders capability transfer. Second, existing speech LLMs struggle to meet the distinct informational needs of understanding and generation. As shown in Figure 1 (a) Left and (b) Left, the Baseline model—trained solely with acoustic tokens—exhibits a contradiction between tasks, i.e., improving one often leads to degradation of the other, highlighting its inability to balance both tasks effectively. In fact, generation tasks demand rich acoustic details (e.g., prosody, emotion, speaker traits) for high-fidelity synthesis (Zeghidour et al. 2021; Défossez et al. 2022; Kumar et al. 2023; Wang et al. 2023), which acoustic tokens capture well but lack high-level semantics (Shi et al. 2024; Dhawan et al. 2024; Anastassiou et al. 2024; Chang et al. 2024). Conversely, understanding tasks benefit from semantic features (Borsos et al. 2023; Rubenstein et al. 2023; Maiti et al. 2024), but semantic tokens inevitably compromise the acoustic details needed for natural speech generation.

To solve these problems, we propose two key insights from the perspectives of speech tokenization and language modeling. First, we present an Understanding-driven Speech Tokenizer (USTokenizer) that can extract high-level semantic features, which are critical for understanding tasks. Unlike prior methods that rely on only self-supervised learning (SSL) representation quantization (Hsu et al. 2021; Chen et al. 2022) or automatic speech recognition (ASR)-based objectives (Du et al. 2024a; Zeng et al. 2024) to capture se-

---

[*] Corresponding authors.

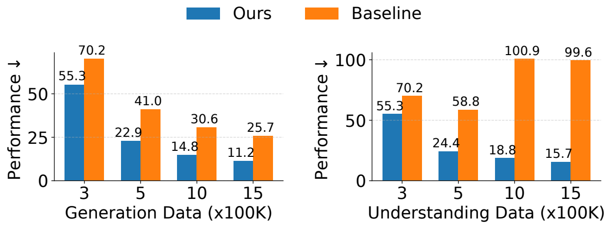
mantics, our approach directly aligns speech tokenizer with the semantic understanding capabilities of text LLMs. This leads to USTokens that have inherently better alignment with text modality, significantly easing modality alignment when adapting text LLMs to speech LLMs. Secondly, building on USTokenizer, we introduce DualSpeechLM, a novel dual speech token modeling framework that effectively handles the distinct informational needs of understanding and generation within a unified end-to-end architecture. Unlike conventional approaches that use the same token as input and output of LLM, our model separates them by using USTokens as input and acoustic tokens as output. Specifically, the USTokenizer provides high-level semantic information to boost understanding tasks, while an AcousticGPT module restores fine-grained acoustic details, enabling diverse and realistic speech generation within a unified end-to-end framework. In addition, we introduce a semantic supervision loss and a Chain of Condition (CoC) strategy to stabilize training and further improve the performance of the unified framework. In summary, our contributions are as follows:

- We present a USTokenizer that can extract high-level semantic information and reduce the modality gap when adapting text LLMs to speech LLMs.
- We propose an end-to-end dual token modeling framework, DualSpeechLM, that simultaneously accepts USTokens as input and generates acoustic tokens as output, effectively accommodating distinct informational needs.
- We also propose a novel semantic supervision loss and a CoC strategy to improve training stability and generation performance.
- Experiments demonstrate that our method achieves faster convergence and excellent performance with small-scale data, and enhances mutual improvements between understanding and generation tasks.

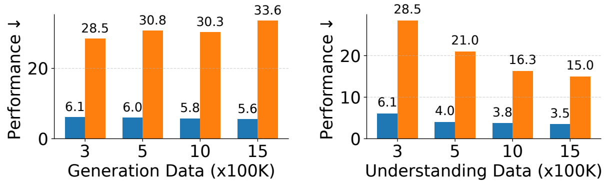
Related Work

Speech Tokenization

The success of autoregressive language models (Achiam et al. 2023; Touvron et al. 2023; Team et al. 2023) has spurred progress in speech LLMs (Chu et al. 2023; Wang et al. 2024c,b; Défossez et al. 2024), where speech tokenizers are essential for converting continuous signals into discrete tokens (Yang et al. 2025b). Speech tokenizers are typically categorized as acoustic or semantic (Borsos et al. 2023; Parker et al. 2024; Yang et al. 2024a). Acoustic tokens, optimized for signal reconstruction (Zeghidour et al. 2021; Défossez et al. 2022; Yang et al. 2023; Wang et al. 2023), capture detailed acoustic features beneficial for generation, but perform poorly on understanding tasks like ASR (Shi et al. 2024; Dhawan et al. 2024; Anastassiou et al. 2024; Chang et al. 2024). Previous semantic tokenizers were trained in two ways: (1) applying clustering (Borsos et al. 2023; Zhang et al. 2023a; Shi et al. 2023) or vector quantization (VQ) (Huang, Meng, and Ko 2024) to representations of SSL models (Hsu et al. 2021; Chen et al. 2022), and (2) applying a VQ layer to the intermediate layer of ASR models (Du et al. 2024a; Zeng et al. 2024; Du et al. 2024b). Although these semantic tokenizers have shown benefits for understanding tasks (Borsos et al. 2023; Rubenstein et al. 2023; Maiti et al. 2024), they do not explicitly consider the alignment to the modality of text LLM (Li et al. 2025). Furthermore, multi-codebook designs (Zhang et al. 2023b; Défossez et al. 2024; Yang et al. 2025b) aim to capture semantics and acoustics in different codebooks jointly, but often introduce complexity when integrated with LLMs. In this work, we present a single VQ-codebook USTokenizer, which not only incorporates high-level semantic information but also achieves better alignment between speech and text modality when applied to LLMs.



(a) Generation performance (↓), measured by generated speech WER (%), using different amounts of generation (left) or understanding (right) data, with the amount of understanding (left) or generation (right) data fixed as 300K samples.



(b) Understanding performance (↓), measured by speech recognition WER (%), using different amounts of generation (left) or understanding (right) data, with the amount of understanding (left) or generation (right) data fixed as 300K samples.

Figure 1: Comparison of baseline and our model on generation and understanding tasks with different ratios of generation and understanding training data.

Speech Language Models

Recent advances in speech LLMs have explored unified understanding and generation (Zhang et al. 2023a; Pan et al. 2024; Défossez et al. 2024; Nguyen et al. 2025). Some speech LLMs (Yang et al. 2024b; Shi et al. 2025) adopt acoustic tokens to ensure high-fidelity speech synthesis. However, such tokens usually lead to degraded performance in understanding tasks, particularly in low-resource scenarios. By contrast, semantic tokens perform better in understanding tasks. Yet, their lack of acoustic details often results in reduced generation quality. To compensate for generation, existing works (Polyak et al. 2021; Zhang et al. 2023a; Du et al. 2024a; Nguyen et al. 2025) introduce additional components such as diffusion models (Ho, Jain, and Abbeel 2020) or flow matching (Lipman et al. 2022), to convert speech tokens into a Mel spectrogram, and then a HiFi-GAN (Kong, Kim, and Bae 2020) vocoder is used to synthesize waveform with the Mel spectrogram as input. These multi-stage pipelines increase complexity and risk error accumulation (Jia et al. 2025). To address these limitations, we propose DualSpeechLM, an end-to-end dual-token modeling framework that explicitly models USTokens as input for understanding and acoustic tokens as output for generation, effectively achieving distinct informational requirements.

Methodologies

We propose a novel unified speech LLM framework, comprising an understanding-driven speech tokenization method and a dual-token modeling paradigm. This framework enhances both understanding and generation capabilities in the resulting DualSpeechLLM. In the following, we describe each module of our framework.

USTokenizer

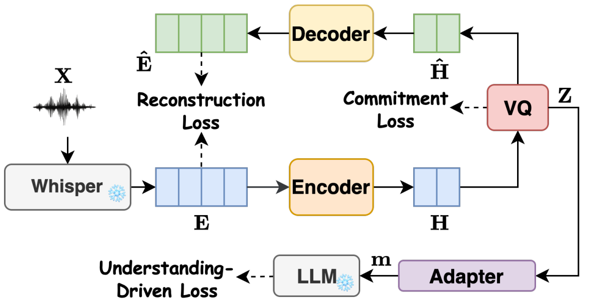


Figure 2: The architecture of USTokenizer, which can extract high-level semantic features aligned with text LLMs via an understanding-driven loss.

As shown in Figure 2, our Understanding-driven Speech Tokenizer (USTokenizer) consists of a pre-trained Whisper encoder, another downsampling Encoder, a vector quantizer (VQ), an upsampling Decoder, an Adapter module, and a frozen text LLM. Given a speech utterance $\mathbf{X}$, we use the final hidden states $\mathbf{E}$ from the Whisper-medium encoder [1] as the input to the downsampling Encoder and then the VQ to obtain the Understanding-driven Speech Token (USToken). Additionally, we novelly project the quantized vector from VQ to the input space of LLM via an adapter to integrate the guidance provided by text LLMs to optimize the speech token during training, as described in the following sections.

Encoder and Decoder The Encoder converts Whisper features $\mathbf{E}$ into representations $\mathbf{H}$ using a 2× downsampling convolution followed by two residual convolutional blocks, then discretized by VQ. Symmetrically, the Decoder reconstructs $\hat{\mathbf{E}}$ from quantized vectors $\hat{\mathbf{H}}$. Both Encoder and Decoder adopt similar residual convolutional blocks with RepCodec (Huang, Meng, and Ko 2024). We then compute the reconstruction loss via mean squared error (MSE) between $\mathbf{E}$ and $\hat{\mathbf{E}}$, formulated as:

$$\mathcal{L}_{\text{reconstruction}}(\mathbf{E},\hat{\mathbf{E}}) = \frac{1}{N}\sum_{i=1}^{N}(E_i - \hat{E}_i)^2, \tag{1}$$

where N is the number of elements in the embedding vector.

Vector Quantization The VQ module discretizes continuous feature representations $\mathbf{H}$ by mapping them to a codebook of learned vectors. In this work, we use a single VQ layer to quantize the feature $\mathbf{H}$ into discrete tokens $\mathbf{Z}$. Following previous works (Huang, Meng, and Ko 2024), we add a commitment loss $\mathcal{L}_{\text{commit}}$ to ensure training stability, with more details provided in *Appendix B*.

Understanding-Driven Loss The LLM module aims to align the speech token $\mathbf{Z}$ with the LLM's input space. By optimizing the understanding tasks upon the text LLM, the required understanding capability will be backpropagated to the optimization of speech token, such that the modality gap between speech and text tokens can be effectively reduced. The LLM-based understanding loss is formulated as the likelihood of generating the target response given a speech prompt using a text LLM (e.g., generating answers given the prompt of speech question):

$$\mathcal{L}_{\text{Under}} = -\sum_{t=1}^{L}\log p(S_t \mid \mathbf{m}, S_{<t};\theta), \tag{2}$$

where L is the length of the sequence, S_t represents the target token at position t, $S_{<t}$ is the sub-sequence of text tokens before t, and $\mathbf{m}$ is the feature of speech prompt extracted by USTokenizer. θ represents parameters of the text LLM.

The LLM module aligns the VQ space with text LLM's input space, which means USTokens are mapped to an LLM-compatible feature space, enhancing subsequent unified understanding and generation modeling. The final total loss of USTokenizer is as follows:

$$\mathcal{L}_{\text{USTokenizer}} = \alpha \cdot \mathcal{L}_{\text{commit}} + \beta \cdot \mathcal{L}_{\text{Under}} + \gamma \cdot \mathcal{L}_{\text{reconstruction}},$$

where α, β and γ are weighting hyperparameter.

DualSpeechLM

As shown in Figure 3, unlike prior Speech LLMs that use the same token for both input and output (Yang et al. 2024b; Wang et al. 2024c; Shi et al. 2025), DualSpeechLM introduces a novel dual-token design by modeling USTokens as input and acoustic tokens as output via an integrated AcousticGPT, effectively accommodating different levels of information required for understanding and generation tasks.

For speech understanding tasks, the USTokenizer first encodes raw speech into USTokens, which are combined with task-specific prompts and fed into the text LLM. The model is trained using Cross-Entropy (CE) loss between predicted and ground-truth text tokens. During inference, the model generates text tokens conditioned on USTokens and prompts. The text tokens are then decoded into the final outputs.

For the generation task, let $\mathbf{U}^{\mathbf{tar}}$ and $\mathbf{A}^{\mathbf{tar}}$ denote USTokens and acoustic tokens of target speech, $\mathbf{U}^{\mathbf{in}}$ and $\mathbf{P}$ represent the USTokens and prompt of input speech. As shown in Figure 3, the text LLM first predicts $\mathbf{U}^{\mathbf{tar}}$ conditioned on $\mathbf{P}$ and $\mathbf{U}^{\mathbf{in}}$, formulated as:

$$p(\mathbf{U}^{\mathbf{tar}}|\mathbf{P},\mathbf{U}^{\mathbf{in}};\theta) = \prod_{t=1}^{L} p({\mathbf{U}^{\mathbf{tar}}}_t|{\mathbf{U}^{\mathbf{tar}}}_{<t},\mathbf{P},\mathbf{U}^{\mathbf{in}};\theta), \tag{3}$$

---

[1] https://github.com/openai/whisper

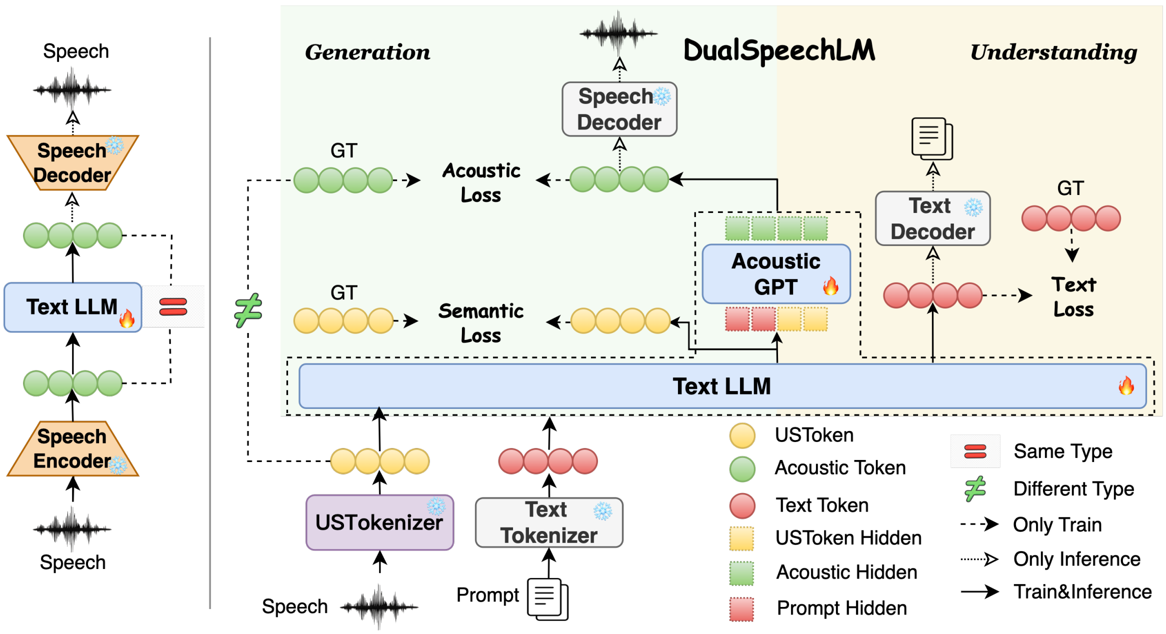


Figure 3: DualSpeechLM's dual-token modeling paradigm. The left illustrates the baseline pipeline treating LLM input/output as identical tokens. In contrast, our DualSpeechLM (right) incorporates an Acoustic GPT module into the text LLM module for joint training, separately processing USToken inputs and acoustic token outputs through distinct modeling paths, effectively capturing the different levels of information required for both generation and understanding tasks.

where L is the length of $\mathbf{U^{tar}}$, ${\mathbf{U^{tar}}}_t$ is the target USToken at position t and ${\mathbf{U^{tar}}}_{<t}$ represents all previous USTokens. θ denotes model parameters. To improve training stability and model performance, we introduce a semantic supervision loss:

$$\mathcal{L}_{\text{semantic}} = -\frac{1}{L}\sum_{t=1}^{L}\log p({\mathbf{U^{tar}}}_t|{\mathbf{U^{tar}}}_{<t}, \mathbf{P}, \mathbf{U^{in}}; \theta) \qquad (4)$$

Next, the predicted USToken sequence $\mathbf{U^{tar}}$ is passed to AcousticGPT, which autoregressively generates acoustic tokens. They are then converted into speech via a Speech Decoder. For both understanding and generation, we adopt text LLM as backbone and fine-tune it using parameter-efficient Low-Rank Adaptation (LoRA) (Hu et al. 2022). Within the unified framework, we integrate AcousticGPT into text LLM to enable joint end-to-end training. This allows DualSpeechLM to simultaneously capture high-level semantics for understanding and fine-grained acoustic details for generation, which we refer to as dual-token modeling. Notably, the acoustic tokens and Speech Decoder are modular and can be implemented using any open-source codec model.

AcousticGPT Inspired by AudioLM (Borsos et al. 2023), we design an AcousticGPT module to autoregressively generate acoustic tokens conditioned on semantic hidden states and speaker embeddings. Built on a GPT-style transformer, the model consists of six causal blocks, each containing a causal self-attention layer followed by a Multi-Layer Perceptron (MLP). The multi-head self-attention captures dependencies under causal constraints. In addition, each block includes residual connections and layer normalization to enhance training stability and model performance. Further details are provided in *Appendix B*.

Chain of Condition To improve training robustness, we introduce a Chain of Condition (CoC) strategy for AcousticGPT. Specifically, given the dynamically changing semantic hidden states, we randomly sample conditioning signals from three sources with equal probability: prompt hidden states $\mathbf{S}_p$, predicted USTokens $\mathbf{U^{tar}}$, or their concatenation $[\mathbf{S}_p; \mathbf{U^{tar}}]$. During inference, we consistently use the concatenated form $[\mathbf{S}_p; \mathbf{U^{tar}}]$. This CoC strategy acts as a regularizer, preventing overfitting to any single modality and improving the model's adaptability by exposing it to varied conditioning signals during training. Importantly, it also mitigates the adverse effects of inaccurate USToken predictions by allowing the model to flexibly rely on more stable prompt hidden states or concatenation. As a result, CoC improves the alignment between USTokens and text tokens in the shared latent space, fostering stronger synergy between understanding and generation representations. The acoustic loss in AcousticGPT is defined as:

$$\mathcal{L}_{\text{acoustic}} = -\sum_{t=1}^{L}\log p({\mathbf{A^{tar}}}_t|{\mathbf{A^{tar}}}_{<t}, \mathbf{S}, \mathbf{Spk}; \theta), \qquad (5)$$

where $\mathbf{S} \in \{\mathbf{S}_p, \mathbf{U^{tar}}, [\mathbf{S}_p; \mathbf{U^{tar}}]\}$ is selected via CoC, $\mathbf{Spk}$ is the speaker embedding, and ${\mathbf{A^{tar}}}_t$ is the predicted acoustic token at time t. $p({\mathbf{A^{tar}}}_t|{\mathbf{A^{tar}}}_{<t}, \mathbf{S}, \mathbf{Spk}; \theta)$ is the predicted probability of generating ${\mathbf{A^{tar}}}_t$, given $\{{\mathbf{A^{tar}}}_{<t}, \mathbf{S}, \mathbf{Spk}\}$. Finally, the overall generation loss combines semantic and acoustic objectives as:

$$\mathcal{L}_{\text{generation}} = \lambda \cdot \mathcal{L}_{\text{semantic}} + \xi \cdot \mathcal{L}_{\text{acoustic}}, \qquad (6)$$

where λ, ξ are weighting hyperparameters.

Model	LLM	Input Token	Training	Data(hrs)	ASR		S2TT	SER	SQA
					Clean	Other	En2De	-	-
					$w \downarrow$		$b4 \uparrow$	$acc \uparrow$	$b4 \uparrow / gs \uparrow$
SpeechGPT (Zhang et al. 2023a)	LLaMA-7B	D(Hubert)	Full FT	70K	42.73	78.54	1.07	-	3.58/40
SpiritLM (Nguyen et al. 2025)	LLaMA-7B	D(Hubert)	Full FT	570K	6.0	11.0	-	-	-
Mini-Omni2 (Xie et al. 2024)	Qwen2-0.5B	C(Whisper)	Full FT	9K	4.8	9.8	-	-	-
VITA (Fu et al. 2025)	Mixtral-8x7B	C(Mel)	Full FT	20K	8.14	18.41	-	-	7.62/70
Moshi (Défossez et al. 2024)	Helium-7B	D(Mimicodec)	Full FT	7M	5.7	-	-	-	-
Qwen2.5-Omni (Xu et al. 2025)	Qwen2.5-7B	C(Whisper)	Full FT	⋆	1.8	3.4	30.2	60.03	4.28/60
Baseline-Acoustic	Phi3.5-3B	D(Wavtokenizer)	LoRA	4.5K	36.52	80.06	1.91	54.90	17.68/76
Baseline-Semantic	Phi3.5-3B	D(Hubert)	LoRA	4.5K	5.70	14.32	11.13	51.91	42.01/85
DualSpeechLM	Phi3.5-3B	D(Hubert)	LoRA	4.5K	5.56	14.62	10.20	51.77	42.59/85
		D(USToken)	LoRA	4.5K	**4.22**	**9.71**	**19.74**	**60.92**	**44.38/88**

Table 1: Understanding capability evaluation. ‘D’: Discrete, ‘C’: Continuous, ‘Full FT’: full fine-tuning. w, $b4$, gs denote WER, BLEU4, ChatGPTScore. *Qwen2.5-Omni uses 300B audio tokens, and DualSpeechLM uses 405M speech tokens.

Model	LLM	Input Token	Training	TTS		VC	T2ST		SC
				Clean	Other	VCTK	Es2En	Fr2En	-
				$s \uparrow / w \downarrow / d \uparrow$		$s \uparrow / w \downarrow / d \uparrow$	$b4 \uparrow$		$b4 \uparrow / gs \uparrow$
GT		-	-	1.0/3.94/3.86	1.0/5.35/3.78	1.0/2.64/3.62	-	-	-/-
SpeechGPT	LLaMA-7B	D(Hubert)	Full FT	-/22.15/3.97	-/24.55/3.96	-	14.62	14.74	3.50/54
Mini-Omni2	Qwen2-0.5B	C(Whisper)	Full FT	-	-	-	-	-	2.22/54
Moshi	Helium-7B	D(Mimicodec)	Full FT	-	-	-	-	-	1.75/50
Qwen2.5-Omni	Qwen2.5-7B	C(Whisper)	Full FT	-/3.73/4.10⋆	-/4.29/4.08	-	-	-	6.70/70
Baseline-Acoustic	Phi3.5-3B	D(Wavtokenizer)	LoRA	0.88/22.11/3.76	0.87/26.38/3.69	0.80/22.07/3.30	8.52	7.62	0.52/62
Baseline-Semantic	Phi3.5-3B	D(Hubert)	LoRA	0.80/21.72/3.29	0.81/22.32/3.26	0.81/18.88/3.25	18.05	15.76	10.44/60
DualSpeechLM	Phi3.5-3B	D(Hubert)	LoRA	0.81/20.80/3.35	0.81/17.12/3.38	**0.82**/18.70/3.37	17.54	15.97	11.06/65
		D(USToken)	LoRA	**0.90/9.25/3.86**	**0.88/9.88/3.82**	0.80/**10.16/3.46**	**26.77**	**23.82**	**16.24/67**

Table 2: Generation capability evaluation. Data usage is the same as **Data(hrs)** in Table 1. s, w, d, $b4$, gs denote SIM, WER, DNSMOS, BLEU4, ChatGPTScore. *Qwen2.5-Omni lacks standalone TTS inference, so the Qwen-TTS API is used instead.

Model	**TTS** ($Q \uparrow / S \uparrow$)	**VC** ($Q \uparrow / S \uparrow$)
GT	$4.20_{\pm 0.33}/3.94_{\pm 0.42}$	$4.48_{\pm 0.23}/4.60_{\pm 0.25}$
SpeechGPT	$3.53_{\pm 0.51}$/-	-
Qwen-TTS	$4.49_{\pm 0.21}$/-	-
Baseline-Acoustic	$3.67_{\pm 0.48}/\mathbf{3.77}_{\pm \mathbf{0.52}}$	$3.63_{\pm 0.33}/3.43_{\pm 0.42}$
Baseline-Semantic	$2.26_{\pm 0.53}/2.76_{\pm 0.45}$	$2.75_{\pm 0.0.56}/2.93_{\pm 0.57}$
Ours-Hubert	$3.48_{\pm 0.61}/3.59_{\pm 0.47}$	$3.54_{\pm 0.34}/3.18_{\pm 0.36}$
Ours-USToken	$\mathbf{3.89}_{\pm \mathbf{0.31}}/3.74_{\pm 0.24}$	$\mathbf{3.80}_{\pm \mathbf{0.28}}/\mathbf{3.55}_{\pm \mathbf{0.37}}$

Table 3: Subjective evaluation on generation tasks. Q, S denote QMOS, SMOS. Ours-Hubert and Ours-USToken refer to DualSpeechLM using Hubert and USToken as input tokens, respectively. Configurations are the same as Table 2.

Experiment Setup

Dataset

We train the USTokenizer using multiple speech understanding tasks, including Automatic Speech Recognition (ASR), Speech Emotion Recognition (SER), and Speech Question Answering (SQA). They are based on LibriSpeech (Panayotov et al. 2015), IEMOCAP (Busso et al. 2008) and SQA dataset from SALMONN (Tang et al. 2023), respectively. In the SQA dataset, both questions and answers were automatically generated based on LibriSpeech transcripts using ChatGPT, forming spoken-question–text-answer pairs.

DualSpeechLM is trained on eight speech understanding and generation tasks using approximately 4,500 hours of speech data. For understanding, we use LibriSpeech for ASR task, the English-to-German (En2De) subset of CoVoST2 (Wang, Wu, and Pino 2020) for Speech-to-Text Translation (S2TT) task, IEMOCAP (Busso et al. 2008) for the SER task, and the SQA dataset (Tang et al. 2023) constructed from LibriSpeech using ChatGPT. For generation, we train Text-to-Speech (TTS) on LibriTTS-R (Koizumi et al. 2023), Text-to-Speech Translation (T2ST) on Spanish-to-English (Es–En) and French-to-English (Fr–En) subsets of CVSS (Jia et al. 2022), and Voice Conversion (VC) on a 1,000-hour subset of LibriHeavy (Kang et al. 2023) with evaluation on 400 pairs from VCTK (Yamagishi et al. 2019). Additionally, the Speech Conversation (SC) task is built by synthesizing speech from the SQA dataset using the Volcengine TTS API. Please see *Appendix C* for dataset details.

Evaluation Metrics

Objective Evaluation For evaluation of the understanding tasks, we use Word Error Rate (WER) for ASR, BLEU-4 for S2TT, accuracy (ACC) for SER, and both BLEU-4 and ChatGPTScore for SQA, respectively. For the generation tasks, we employ speaker similarity (SIM), WER, and

Model	LLM	Token	ASR		S2TT	SER	SQA
			Clean	Other	En2De	-	-
			$wer \downarrow$		$bleu4 \uparrow$	$acc \uparrow$	$bleu4 \uparrow$
Ours	Phi3.5-3B	Hubert	5.56	14.62	10.20	51.77	42.59
		USToken	4.22	9.71	19.74	**60.92**	44.38
Ours	Vicuna-7B	Hubert	5.28	16.64	10.92	54.08	42.68
		USToken	**4.15**	**9.69**	**20.46**	58.42	**44.60**

Table 4: The understanding performance comparison of DualSpeechLM when using different text LLM backbones.

DNSMOS to assess TTS and VC performance. BLEU-4 is used for T2ST. Both BLEU-4 and ChatGPTScore are used for evaluating the SC task.

Subjective Evaluation We also conduct subjective evaluations for generation tasks, focusing primarily on TTS and VC. For each task, subjective assessment is carried out from speech quality (QMOS) and speaker similarity (SMOS). More details can be found in *Appendix C*.

Model Settings

In experiments, we compare four variants under the same configurations: (1) Baseline-Acoustic, where both input and output are acoustic token from WavTokenizer (Ji et al. 2024); (2) Baseline-Semantic, where both input and output are HuBERT token (Hsu et al. 2021); (3) DualSpeechLM-Hubert, using HuBERT token as input, acoustic token from WavTokenizer as output; and (4) DualSpeechLM-USToken, using our USToken as input, acoustic token from WavTokenizer as output. Notably, the Baseline-Acoustic model requires 160k steps to converge under multi-task settings, while others converge in 60k steps. Further details are provided in the *Appendix D*.

Results and Analyses

Understanding and Generation Performance

We first compare the performance of our method with prior work. The results for understanding and generation are shown in Table 1, 2 and 3, respectively. From these results, we observe the following:

(1) *USToken has better modality commonality with text, reducing the difficulty of modality alignment when adapting text LLMs to speech LLMs.* DualSpeechLM-USToken outperforms both Baselines and DualSpeechLM-Hubert across almost all tasks, highlighting USToken's enhanced semantic understanding capabilities. Additionally, on translation tasks (S2TT and T2ST), which require the inherent translation capabilities of text LLM, our DualSpeechLM-USToken still significantly outperforms both Baselines and DualSpeechLM-Hubert. This demonstrates that USToken has better alignment with text modality, allowing it to retain more capabilities of text LLM during training. Overall, both DualSpeechLM-Hubert and DualSpeechLM-USToken exhibit better performance than the baseline, indicating that the design of Acoustic GPT in DualSpeechLM also helps alleviate the pressure of text LLM. The comparison of convergence speed is provided in *Appendix A*.

(2) *DualSpeechLM achieves strong performance in both tasks with small-scale resources.* Specifically, compared to Baselines, both DualSpeechLM-Hubert and DualSpeechLM-USToken demonstrate significant performance improvements, highlighting the effectiveness of DualSpeechLM. Compared to existing unified models, which rely on larger datasets and full fine-tuning, such as SpeechGPT and Mini-Omni2, DualSpeechLM-USToken achieves better performance with just 4.5k hours of data using parameter-efficient fine-tuning, LoRA.

(3) *DualSpeechLM simultaneously fulfill the distinct information requirements of generation and understanding by dual-token modeling.* Comparisons between Baseline-Acoustic and Baseline-Semantic in Table 1, 2, and 3 show that semantic tokens excel in understanding tasks, while acoustic tokens perform better in generation, e.g., higher SIM and DNSMOS scores in TTS. However, acoustic tokens carry weaker semantics reflected by higher WER results and lower translation quality. In contrast, our DualSpeechLM demonstrates superior performance on both understanding and generation tasks simultaneously due to its dual-token modeling. Another evidence is shown in Figure 1, where our method consistently improves performance on both generation and understanding tasks as the amount of either type of training data increases. This demonstrates its strong ability to support one task without compromising the other.

Performance with Different LLM Backbones

To further assess the generalizability of DualSpeechLM and the effectiveness of USToken across different backbones, we compare models using various text LLMs as backbones. Table 4 and 5 show that DualSpeechLM-USToken consistently outperforms DualSpeechLM-Hubert across almost all tasks, regardless of the underlying LLMs. This demonstrates that USToken not only aligns more effectively with text LLMs but also transfers well across model architectures. Additionally, our DualSpeechLM achieves performance comparable to or better than existing methods, with different text LLMs as the backbone. This highlights the strong compatibility and robustness of DualSpeechLM across different LLM backbones, further demonstrating its effectiveness as a unified framework for both speech understanding and generation.

Ablation Study

We perform ablation studies to evaluate the contribution of each component, with results shown in Table 6 and 7.

The Influence of Understanding-driven Loss. We ablate the understanding-driven loss in USTokenizer while keeping all other components unchanged. As shown in Table 7, removing this LLM-based objective leads to significant performance drops in understanding tasks (e.g., BLEU4 decreases from 44.38 to 37.67 on SQA), highlighting its importance for aligning USTokens more closely with the semantic space of text LLMs using the understanding-driven loss. For T2ST and SC in Table 6, we also observe a performance drop after removing the understanding-driven loss. This can be attributed to these two tasks rely more heavily on the internal capabilities of the text LLM. The understanding-driven

Model	LLM	Input Token	TTS		VC	T2ST		SC
			Clean	Other	VCTK	Es2En	Fr2En	-
			$s\uparrow/w\downarrow/d\uparrow$		$s\uparrow/w\downarrow/d\uparrow$	$bleu4\uparrow$		$bleu4\uparrow$
DualSpeechLM	Phi3.5-3B	Hubert	0.81/20.80/3.35	0.81/17.12/3.38	**0.82**/18.70/3.37	17.54	15.97	11.06
		USToken	**0.90/9.25**/3.86	**0.88/9.88**/3.82	0.80/**10.16/3.46**	26.77	23.82	16.24
DualSpeechLM	Vicuna-7B	Hubert	0.87/13.58/3.88	0.86/15.56/3.80	0.78/19.86/3.38	22.29	19.74	13.17
		USToken	0.89/12.63/**3.90**	0.87/14.65/**3.86**	0.78/17.03/3.43	**29.58**	**25.40**	**17.18**

Table 5: Comparison of generation performance of DualSpeechLM using different text LLM backbones.

Model	TTS		VC	T2ST		SC
	Clean	Other	VCTK	Es2En	Fr2En	-
	$s\uparrow/w\downarrow/d\uparrow$		$s\uparrow/w\downarrow/d\uparrow$	$bleu4\uparrow$		$bleu4\uparrow$
DualSpeechLM	0.90/9.25/3.86	0.88/9.88/3.82	0.80/10.16/3.46	26.77	23.82	16.24
USTokenizer Ablation						
w/o Understanding-driven Loss	0.90/8.59/3.86	0.88/9.45/3.81	0.82/9.90/3.39	24.03	22.73	15.59
w/o Reconstruction Loss	0.83/52.50/3.02	0.81/54.24/2.96	0.79/26.08/3.33	32.43	29.46	17.09
DualSpeechLM Ablation						
w/o Semantic Loss	0.80/167.56/3.85	0.85/175.11/3.83	0.80/264.53/3.45	0.09	0.07	0.15
w/o CoC	0.89/9.96/3.84	0.87/10.34/3.81	0.79/13.06/3.38	-	-	-

Table 6: Ablation study using generation tasks. We use Phi-3.5-3B as the text LLM backbone.

Model	ASR		S2TT	SER	SQA
	Clean	Other	En2De	-	-
	$wer\downarrow$		$bleu4\uparrow$	$acc\uparrow$	$bleu4\uparrow$
DualSpeechLM	4.22	9.71	19.74	60.92	44.38
USTokenizer Ablation					
w/o Understanding-driven Loss	4.81	10.43	14.81	52.7	37.67
w/o Reconstruction Loss	4.74	10.46	18.5	45.21	46.44
DualSpeechLM Ablation					
w/o Semantic Loss	4.31	9.61	18.67	60.35	43.86

Table 7: Ablation study using understanding tasks. Phi-3.5-3B is used as the text LLM backbone.

loss plays a critical role in enhancing modality commonality between USTokens and text tokens, thereby helping to preserve as much of the original ability of the text LLM as possible when adapting the text LLM to speech-related tasks.

Effect of the Reconstruction Loss. The third row in Table 6 and 7 shows the results when the reconstruction loss is removed from USTokenizer. For tasks like T2ST that rely more on the reasoning capabilities of the text LLM, performance actually improves. This suggests that removing the reconstruction objective pushes USTokenizer to better align with the text modality, thus narrowing the modality gap and allowing the LLM to operate more effectively on semantically rich tasks. However, we observe notable performance degradation in SER, TTS, and VC tasks, indicating that the reconstruction loss is essential for preserving fine-grained information, which is crucial for these tasks.

Effect of Semantic Loss. The fourth row in Table 6 and 7 shows the results of removing semantic loss from DualSpeechLM. This leads to a slight decline in understanding performance, indicating that reconstructing USTokens, as enforced by semantic loss, provides useful semantic supervision that benefits understanding tasks. More critically, we observe a substantial performance drop across all generation tasks. Without semantic loss, the text LLM struggles with producing high-quality USTokens, resulting in degraded inputs to AcousticGPT and ultimately poor generation of acoustic tokens. These results underscore the pivotal role of semantic loss in ensuring accurate semantic representation, which is essential not only for understanding but also for maintaining high-fidelity generation in unified DualSpeechLM.

Influence of CoC. The final row in Table 6 shows the impact of removing the Chain of Condition (CoC) strategy from the AcousticGPT module, which is applied only to TTS and VC tasks. Performance drops notably on both tasks, indicating that CoC's stochastic conditioning improves alignment between DualTokens and text tokens in the shared latent space. This enhanced alignment leads to more stable and accurate acoustic token generation.

Conclusions

In this work, we aim to develop a unified speech LLM that can simultaneously excel in both speech understanding and generation. We first propose USTokenizer, which reduces the modality gap between speech and text when adapting text LLMs to speech LLMs. Based on this, we present an end-to-end DualSpeechLM that effectively accommodates the different informational requirements for understanding and generation by a dual-token modeling strategy. Experiments indicate that our method achieves significantly better performance than baselines, enabling mutual improvements between understanding and generation. In the future, we plan to improve DualSpeechLM to larger, more diverse multilingual and cross-domain datasets to further explore its generalization and robustness.

Acknowledgements

This work is partially supported by the General Research Fund from the Research Grants Council of Hong Kong SAR Government (Project No. 14202623), National Natural Science Foundation of China (62076144), Shenzhen Science and Technology Program (JCYJ20220818101014030), and Ant Group.

References

Achiam, J.; Adler, S.; Agarwal, S.; et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Anastassiou, P.; Chen, J.; Chen, J.; Chen, Y.; Chen, Z.; Chen, Z.; Cong, J.; Deng, L.; Ding, C.; Gao, L.; et al. 2024. Seed-tts: A family of high-quality versatile speech generation models. *arXiv preprint arXiv:2406.02430*.

Borsos, Z.; Marinier, R.; Vincent, D.; Kharitonov, E.; Pietquin, O.; Sharifi, M.; Roblek, D.; Teboul, O.; Grangier, D.; Tagliasacchi, M.; et al. 2023. Audiolm: a language modeling approach to audio generation. *IEEE/ACM transactions on audio, speech, and language processing*, 31: 2523–2533.

Busso, C.; Bulut, M.; Lee, C.-C.; et al. 2008. IEMOCAP: Interactive emotional dyadic motion capture database. *Language resources and evaluation*, 42: 335–359.

Chang, X.; Yan, B.; Choi, K.; et al. 2024. Exploring Speech Recognition, Translation, and Understanding with Discrete Speech Units: A Comparative Study. In *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 11481–11485.

Chen, S.; Wang, C.; Chen, Z.; Wu, Y.; Liu, S.; Chen, Z.; Li, J.; Kanda, N.; Yoshioka, T.; Xiao, X.; et al. 2022. Wavlm: Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing*, 16(6): 1505–1518.

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.

Défossez, A.; Copet, J.; Synnaeve, G.; and Adi, Y. 2022. High fidelity neural audio compression. *arXiv preprint arXiv:2210.13438*.

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

Dhawan, K.; Koluguri, N. R.; Jukić, A.; Langman, R.; Balam, J.; and Ginsburg, B. 2024. Codec-ASR: Training Performant Automatic Speech Recognition Systems with Discrete Speech Representations. In *Proc. Interspeech 2024*, 2574–2578.

Du, Z.; Chen, Q.; Zhang, S.; Hu, K.; Lu, H.; Yang, Y.; Hu, H.; Zheng, S.; Gu, Y.; Ma, Z.; et al. 2024a. Cosyvoice: A scalable multilingual zero-shot text-to-speech synthesizer based on supervised semantic tokens. *arXiv preprint arXiv:2407.05407*.

Du, Z.; Wang, Y.; Chen, Q.; et al. 2024b. CosyVoice 2: Scalable Streaming Speech Synthesis with Large Language Models. arXiv:2412.10117.

Fu, C.; Lin, H.; Long, Z.; Shen, Y.; Dai, Y.; Zhao, M.; Zhang, Y.-F.; Dong, S.; et al. 2025. VITA: Towards Open-Source Interactive Omni Multimodal LLM. arXiv:2408.05211.

Gong, Y.; Liu, A. H.; Luo, H.; Karlinsky, L.; and Glass, J. 2023. Joint audio and speech understanding. In *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 1–8. IEEE.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33: 6840–6851.

Hsu, W.-N.; Bolte, B.; Tsai, Y.-H. H.; et al. 2021. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM transactions on audio, speech, and language processing*, 29: 3451–3460.

Hu, E. J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; Chen, W.; et al. 2022. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2): 3.

Huang, Z.; Meng, C.; and Ko, T. 2024. RepCodec: A Speech Representation Codec for Speech Tokenization. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 5777–5790.

Ji, S.; Jiang, Z.; Wang, W.; Chen, Y.; Fang, M.; Zuo, J.; Yang, Q.; Cheng, X.; Wang, Z.; Li, R.; et al. 2024. Wavtokenizer: an efficient acoustic discrete codec tokenizer for audio language modeling. *arXiv preprint arXiv:2408.16532*.

Jia, D.; Chen, Z.; Chen, J.; et al. 2025. DiTAR: Diffusion Transformer Autoregressive Modeling for Speech Generation. *arXiv preprint arXiv:2502.03930*.

Jia, Y.; Tadmor Ramanovich, M.; Wang, Q.; and Zen, H. 2022. CVSS Corpus and Massively Multilingual Speech-to-Speech Translation. In *Proceedings of Language Resources and Evaluation Conference (LREC)*, 6691–6703.

Kang, W.; Yang, X.; Yao, Z.; et al. 2023. Libriheavy: a 50,000 hours ASR corpus with punctuation casing and context. arXiv:2309.08105.

Kim, J.; Lee, K.; Chung, S.; and Cho, J. 2024. Clam-tts: Improving neural codec language model for zero-shot text-to-speech. *arXiv preprint arXiv:2404.02781*.

Koizumi, Y.; Zen, H.; Karita, S.; Ding, Y.; Yatabe, K.; Morioka, N.; Bacchiani, M.; Zhang, Y.; Han, W.; and Bapna, A. 2023. LibriTTS-R: A Restored Multi-Speaker Text-to-Speech Corpus. arXiv:2305.18802.

Kong, J.; Kim, J.; and Bae, J. 2020. Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis. *Advances in neural information processing systems*, 33: 17022–17033.

Kumar, R.; Seetharaman, P.; Luebs, A.; Kumar, I.; and Kumar, K. 2023. High-fidelity audio compression with improved rvqgan. *Advances in Neural Information Processing Systems*, 36: 27980–27993.

Li, T.; Liu, J.; Zhang, T.; et al. 2025. Baichuan-Audio: A Unified Framework for End-to-End Speech Interaction. arXiv:2502.17239.

Lipman, Y.; Chen, R. T.; Ben-Hamu, H.; Nickel, M.; and Le, M. 2022. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*.

Maiti, S.; Peng, Y.; Choi, S.; et al. 2024. Voxtlm: Unified decoder-only models for consolidating speech recognition, synthesis and speech, text continuation tasks. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 13326–13330. IEEE.

Nguyen, T. A.; Muller, B.; Yu, B.; et al. 2025. Spirit-lm: Interleaved spoken and written language model. *Transactions of the Association for Computational Linguistics*, 13: 30–52.

Pan, K.; Tang, S.; Li, J.; et al. 2024. Auto-encoding morph-tokens for multimodal LLM. In *Proceedings of the 41st International Conference on Machine Learning*, 39308–39323.

Panayotov, V.; Chen, G.; Povey, D.; and Khudanpur, S. 2015. Librispeech: An ASR corpus based on public domain audio books. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 5206–5210.

Parker, J. D.; Smirnov, A.; Pons, J.; et al. 2024. Scaling transformers for low-bitrate high-quality speech coding. *arXiv preprint arXiv:2411.19842*.

Polyak, A.; Adi, Y.; Copet, J.; et al. 2021. Speech Resynthesis from Discrete Disentangled Self-Supervised Representations. In *INTERSPEECH 2021-Annual Conference of the International Speech Communication Association*.

Rubenstein, P. K.; Asawaroengchai, C.; Nguyen, D. D.; et al. 2023. Audiopalm: A large language model that can speak and listen. *arXiv preprint arXiv:2306.12925*.

Shi, J.; Inaguma, H.; Ma, X.; et al. 2023. Multi-resolution HuBERT: Multi-resolution speech self-supervised learning with masked unit prediction. *arXiv preprint arXiv:2310.02720*.

Shi, J.; Tian, J.; Wu, Y.; et al. 2024. Espnet-codec: Comprehensive training and evaluation of neural codecs for audio, music, and speech. In *2024 IEEE Spoken Language Technology Workshop (SLT)*, 562–569. IEEE.

Shi, J.; Zhang, C.; Tian, J.; Ni, J.; Zhang, H.; Watanabe, S.; and Yu, D. 2025. Balancing Speech Understanding and Generation Using Continual Pre-training for Codec-based Speech LLM. *arXiv preprint arXiv:2502.16897*.

Tang, C.; Yu, W.; Sun, G.; et al. 2023. Salmonn: Towards generic hearing abilities for large language models. *arXiv preprint arXiv:2310.13289*.

Team, G.; Anil, R.; Borgeaud, S.; et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Touvron, H.; Lavril, T.; Izacard, G.; et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Wang, C.; Chen, S.; Wu, Y.; et al. 2023. Neural codec language models are zero-shot text to speech synthesizers. *arXiv preprint arXiv:2301.02111*.

Wang, C.; Liao, M.; Huang, Z.; Lu, J.; Wu, J.; Liu, Y.; Zong, C.; and Zhang, J. 2024a. BLSP: Bootstrapping Language-Speech Pre-training via Behavior Alignment of Continuation Writing. arXiv:2309.00916.

Wang, C.; Wu, A.; and Pino, J. 2020. CoVoST 2: A Massively Multilingual Speech-to-Text Translation Corpus. arXiv:2007.10310.

Wang, X.; Li, Y.; Fu, C.; et al. 2024b. Freeze-omni: A smart and low latency speech-to-speech dialogue model with frozen llm. *arXiv preprint arXiv:2411.00774*.

Wang, X.; Thakker, M.; Chen, Z.; et al. 2024c. Speechx: Neural codec language model as a versatile speech transformer. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*.

Wang, Y.; Chen, H.; Yang, D.; et al. 2025. UniSep: Universal Target Audio Separation with Language Models at Scale. *arXiv preprint arXiv:2503.23762*.

Xie, Z.; et al. 2024. Mini-Omni2: Towards Open-source GPT-4o with Vision, Speech and Duplex Capabilities. arXiv:2410.11190.

Xu, J.; Guo, Z.; He, J.; et al. 2025. Qwen2. 5-omni technical report. *arXiv preprint arXiv:2503.20215*.

Yamagishi, J.; et al. 2019. CSTR VCTK Corpus: English Multi-speaker Corpus for CSTR Voice Cloning Toolkit (version 0.92).

Yang, D.; Guo, H.; Wang, Y.; et al. 2024a. UniAudio 1.5: large language model-driven audio codec is a few-shot audio task learner. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, 56802–56827.

Yang, D.; Huang, R.; Wang, Y.; et al. 2025a. Simplespeech 2: Towards simple and efficient text-to-speech with flow-based scalar latent transformer diffusion models. *IEEE Transactions on Audio, Speech and Language Processing*.

Yang, D.; Liu, S.; Guo, H.; et al. 2025b. ALMTokenizer: A Low-bitrate and Semantic-rich Audio Codec Tokenizer for Audio Language Modeling. In *Forty-second International Conference on Machine Learning*.

Yang, D.; Liu, S.; Huang, R.; et al. 2023. Hifi-codec: Group-residual vector quantization for high fidelity audio codec. *arXiv preprint arXiv:2305.02765*.

Yang, D.; Tian, J.; Tan, X.; et al. 2024b. Uniaudio: Towards universal audio generation with large language models. In *Forty-first International Conference on Machine Learning*.

Zeghidour, N.; Luebs, A.; Omran, A.; Skoglund, J.; and Tagliasacchi, M. 2021. Soundstream: An end-to-end neural audio codec. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30: 495–507.

Zeng, A.; Du, Z.; Liu, M.; et al. 2024. GLM-4-Voice: Towards Intelligent and Human-Like End-to-End Spoken Chatbot. arXiv:2412.02612.

Zhang, D.; Li, S.; Zhang, X.; et al. 2023a. Speechgpt: Empowering large language models with intrinsic cross-modal conversational abilities. *arXiv preprint arXiv:2305.11000*.

Zhang, X.; Zhang, D.; Li, S.; et al. 2023b. SpeechTokenizer: Unified Speech Tokenizer for Speech Language Models. arXiv:2308.16692.