DialectS2S: End-to-End Speech Dialogue Modeling for Low-Resource Chinese Dialects

Yi Shu[1], Tianyu Peng[1,2,3], Yingzhuo Deng[1,2], Wen Yang[1,2], Jun Lin[4], Changming Xie[4], Xinyu Yu[4], and Jiajun Zhang[1,2,3(✉)]

[1] School of Artificial Intelligence, University of Chinese Academy of Sciences
[2] Institute of Automation, Chinese Academy of Sciences
[3] Wuhan AI Research
[4] GWM AI Lab
shuyi23@mails.ucas.ac.cn jjzhang@nlpr.ia.ac.cn

Abstract. Current end-to-end speech dialogue models are primarily optimized for mainstream languages and remain limited in low-resource dialect scenarios due to the scarcity of dialect speech data. Moreover, during dialect adaptation, the semantic representation space of speech dialogue models continuously evolves, while conventional speech supervision remains unchanged, leading to semantic inconsistency between hidden representations and speech targets and degrading speech stability and naturalness. To address these issues, we propose DialectS2S, an end-to-end speech dialogue model for Chinese dialects. We first develop a scalable dialect speech dialogue synthesis pipeline for efficient data construction. We further introduce a two-stage post-training strategy with self-aligned speech supervision, which aligns the semantic content of speech supervision with the evolved semantic representations of the model to improve dialect speech generation quality. Experimental results show that DialectS2S consistently outperforms existing baselines across multiple Chinese dialects in speech dialogue, achieving substantial improvements in dialect consistency, response quality, and speech intelligibility. Our work provides an efficient and scalable solution for end-to-end speech dialogue modeling in low-resource dialect scenarios. To facilitate future research and practical applications, we fully open-source the DialectS2S framework, including model checkpoints, training datasets, and fine-tuning code.

Keywords: Speech-to-Speech Dialogue · Chinese Dialects · Self-aligned Speech Supervision · End-to-End Large Speech Language Models

1 Introduction

Dialects play an important role in regional communication and cultural expression. Recent end-to-end speech dialogue models [23,19,6,18,17,3] have achieved

(✉) Corresponding author
Resources: model checkpoints, training datasets, and fine-tuning code.

strong conversational performance in mainstream languages such as Mandarin Chinese and English. However, support for low-resource Chinese dialects remains limited. Existing Chinese dialect speech research mainly focuses on speech recognition or speech synthesis, while unified end-to-end speech dialogue modeling for Chinese dialects remains largely underexplored.

Current speech dialogue models often rely on large-scale speech dialogue corpora, while high-quality dialect speech data remain scarce. Moreover, recent studies have shown that directly supervised fine-tuning speech dialogue models may degrade speech quality and generation stability [24]. This issue becomes more challenging in low-resource dialect scenarios due to limited high-quality supervision and large acoustic variations. Therefore, enabling stable and natural dialect speech interaction under limited supervision remains a major challenge for end-to-end speech dialogue systems.

To address these challenges, we propose DialectS2S, an end-to-end speech dialogue model for low-resource Chinese dialects. DialectS2S supports multilingual interaction across five languages and dialects, including Mandarin, English, Sichuanese, Cantonese, and the Tianjin dialect. To construct dialect speech dialogue data, we develop a scalable synthesis pipeline that rewrites existing dialogue texts into dialectal expressions using large language models and synthesizes corresponding dialect speech with speech generation models. Furthermore, we introduce a two-stage post-training strategy with self-aligned speech supervision to improve speech generation stability and naturalness during dialect adaptation.

In summary, our main contributions are as follows:

(1) **End-to-End Speech Dialogue Model for Chinese Dialects:** We propose DialectS2S, an end-to-end speech dialogue model for low-resource Chinese dialects. To facilitate future research on dialect speech interaction, we fully open-source the DialectS2S framework, including model checkpoints, training datasets, and fine-tuning code.
(2) **Self-Aligned Speech Supervision for Dialect Speech Generation:** We propose a two-stage post-training strategy with self-aligned speech supervision, which aligns speech targets with the model's semantic predictions and improves speech intelligibility while preserving response quality.
(3) **Strong Multi-Dialect Performance:** Experimental results demonstrate that DialectS2S significantly outperforms existing open-source baselines in dialect consistency, response quality, and speech intelligibility across multiple Chinese dialects.

2 Related Work

2.1 End-to-End Large Speech Models

Recent end-to-end large speech models have achieved strong performance in speech recognition, synthesis, and dialogue tasks. Compared with cascaded systems, these models directly map speech inputs to text and speech outputs, reducing error accumulation and improving interaction naturalness. LLaMA-Omni [8]

appends a speech decoder to a large language model for unified speech understanding and generation. GLM-4-Voice [23] adopts interleaved text-speech modeling for streaming interaction. Step-Audio 2 [18] and Step-Audio-R1 [16] follow a similar interleaved architecture. Moshi [5], Voila [14] and MiniCPM-o [3] further support low-latency full-duplex conversations. More recently, Qwen2.5-Omni [19] and Qwen3-Omni [20] adopt the Thinker-Talker architecture, where a large language model handles semantic reasoning and a smaller model generates speech. OpenS2S [17] further improves fine-grained speech understanding and releases open-source models and training pipelines. Despite these advances, existing systems are mainly optimized for high-resource languages such as Mandarin and English, while low-resource dialect speech interaction remains underexplored.

2.2 Low-Resource Dialect Speech Modeling

Compared with high-resource languages such as English and Mandarin Chinese, speech modeling for Chinese dialect varieties faces challenges including limited data availability, expensive annotation, and substantial pronunciation variation. Existing studies mainly focus on dialect speech corpora, automatic speech recognition (ASR), and text-to-speech synthesis (TTS). KeSpeech [15] constructs a large-scale corpus covering eight Chinese dialects and more than 27,000 speakers, while WenetSpeech-Yue [9] and WenetSpeech-Chuan [4] provide high-quality Cantonese and Sichuanese speech datasets. Systems such as FunASR [1] and FireRedASR2S [21] support multilingual and multi-dialect speech processing. In TTS, DIAMOE-TTS [2] enables unified dialect modeling through IPA representations, while CosyVoice2 [7] improves speech naturalness and speaker consistency and supports dialect speech generation from reference audio. However, existing work primarily addresses speech recognition or synthesis in isolation, with limited exploration of end-to-end spoken dialogue systems for Chinese dialects.

3 Method

3.1 Overview

We propose an efficient framework for low-resource dialect adaptation in end-to-end speech dialogue models. The framework consists of a dialect speech dialogue synthesis pipeline and a two-stage post-training procedure, including mixed-data supervised fine-tuning and self-aligned speech supervision training. Through the proposed framework, pretrained speech dialogue models can acquire stable dialect understanding and generation capabilities while preserving their original multilingual ability.

3.2 Dialect Speech Dialogue Synthesis Pipeline

The proposed synthesis pipeline consists of three stages: dialect text generation, dialect speech synthesis, and speech naturalness filtering.

Dialect Dialogue Generation Directly prompting large language models to generate dialect dialogues often produces repetitive and low-diversity content. Instead, we rewrite the dialogue portion of an open-source Mandarin dialogue dataset [17] into dialectal expressions using large language models [22] while preserving the original semantics and dialogue structure. This strategy efficiently reuses existing dialogue corpora and enables low-cost construction of diverse dialect dialogue texts.

Dialect Speech Synthesis We adopt different synthesis strategies for user-query speech and system-response speech. For user-query speech, we select 100 dialect seed utterances for each dialect from open-source dialect datasets [15,9,4], including both male and female speakers. These utterances are used as reference audio for CosyVoice2 [7] to synthesize dialect query speech from the rewritten dialogue texts, ensuring diverse vocal characteristics in the input speech. For system-response speech, we adopt a consistent speaker timbre strategy for response synthesis. Specifically, we first construct dialect-specific reference speech conditioned on a target speaker and then use the generated reference speech for response synthesis. This strategy maintains consistent voice characteristics in synthesized responses and further improves dialect expressiveness and speech naturalness. Further seed-set details are provided in Appendix A.

Speech Naturalness Filtering To improve data quality, we further introduce a speech filtering stage using UTMOS [13] to automatically remove synthesized samples with unclear pronunciation or unnatural prosody. Filtering details are provided in Appendix A.

3.3 Model Training

Most existing end-to-end speech-to-speech dialogue models [23,19,6,18,17,20,16] are built upon pretrained text language models and synchronously generate both text and speech responses from speech inputs. In this work, we interpret such systems from a unified Thinker-Talker perspective. Specifically, given an input speech sequence $\mathbf{X}$, the *Thinker* module performs speech understanding and semantic reasoning, producing intermediate hidden representations $\mathbf{H}$ together with text response tokens $\mathbf{T}$. Conditioned on $\mathbf{H}$, the *Talker* module learns the mapping from hidden semantic representations to speech response tokens $\mathbf{U}$ by leveraging the semantic and paralinguistic information encoded in the hidden states. The overall architecture is illustrated in Figure 1.

Although these models effectively leverage pretrained language models for end-to-end speech interaction, adapting them to low-resource dialect scenarios often introduces shifts in the hidden-state distribution during post-training. Such shifts are not necessarily undesirable, since the model must gradually acquire new dialect-specific semantic and acoustic representations. However, the speech supervision used during training often remains semantically inconsistent with the evolved hidden representations. As a result, the Talker module must implicitly perform semantic recovery in addition to speech generation, increasing optimization difficulty and degrading speech quality.

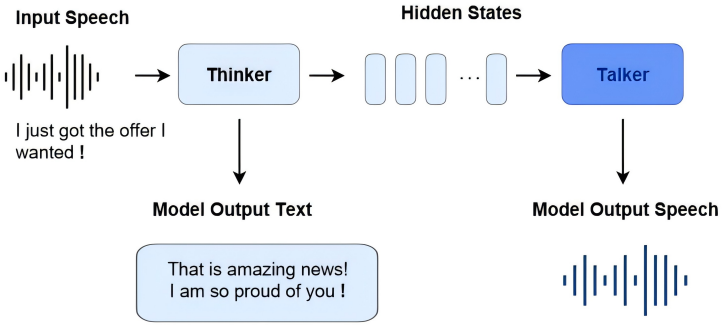

Fig. 1. The proposed unified Thinker-Talker architecture.

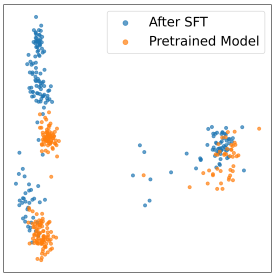

Fig. 2. Visualization of the final-layer Thinker hidden states before and after supervised fine-tuning.

As illustrated in Figure 2, we visualize the hidden-state distributions using PCA projection on the token-averaged final-layer Thinker representations. The hidden representations continuously evolve after supervised fine-tuning, highlighting the necessity of adapting speech supervision to the updated semantic space.

To better align speech supervision with the evolving semantic representations during dialect adaptation, we adopt a two-stage training framework consisting of **Mixed-Data SFT** and **Self-aligned Speech Supervision Training**. The first stage enables the model to acquire dialect-aware semantic representations through supervised fine-tuning, while the second stage further aligns speech generation with the evolved hidden representations to improve dialect speech quality and intelligibility.

Mixed-Data SFT Initialized from a speech dialogue model [17] with strong conversational performance in mainstream languages, we perform full-parameter supervised fine-tuning on mixed dialect, Mandarin, and English speech dialogue data. The training objective is defined as:

$$\mathcal{L}_{SFT} = \underbrace{-\sum_{i=1}^{L} \log P(t_i \mid \mathbf{X}, \mathbf{T}_{<i}; \theta_{Thinker})}_{\text{Thinker Loss}} \underbrace{-\sum_{j=1}^{V} \log P(u_j \mid \mathbf{H}, \mathbf{U}_{<j}; \theta_{Talker})}_{\text{Talker Loss}} \quad (1)$$

where $\theta_{Thinker}$ and θ_{Talker} denote the parameters of the Thinker and Talker modules.

This stage gradually shifts the hidden representations of the Thinker module toward dialect-aware semantic spaces through supervised fine-tuning, enabling the model to acquire initial dialect interaction capabilities. However, the Talker

module still struggles to fully adapt to the evolved semantic representations, since the representation space of the Thinker module continues to evolve during training while the speech supervision remains unchanged.

Self-aligned Speech Supervision Training To better align speech supervision with the evolved semantic representations of the Thinker module, we further propose self-aligned speech supervision training.

We first use the fine-tuned Thinker module to generate text predictions:

$$\hat{\mathbf{T}} = \text{Thinker}(\mathbf{X}) \tag{2}$$

Aligned speech supervision is then synthesized from the predicted text using a dialect TTS model:

$$\hat{\mathbf{U}} = \text{TTS}(\hat{\mathbf{T}}) \tag{3}$$

The synthesized speech is subsequently used as the new supervision target, such that the semantic content of the speech supervision is explicitly aligned with the semantic information encoded in the hidden representations for Talker modeling. To maintain stable coordination between the Thinker and Talker modules, the self-aligned stage still adopts end-to-end full-parameter optimization, although the self-aligned loss is primarily applied to the Talker module.

$$\mathcal{L}_{Align\_Talker} = -\sum_{j=1}^{V} \log P(\hat{u}_j \mid \mathbf{H}, \hat{\mathbf{U}}_{<j}; \theta_{Talker}) \tag{4}$$

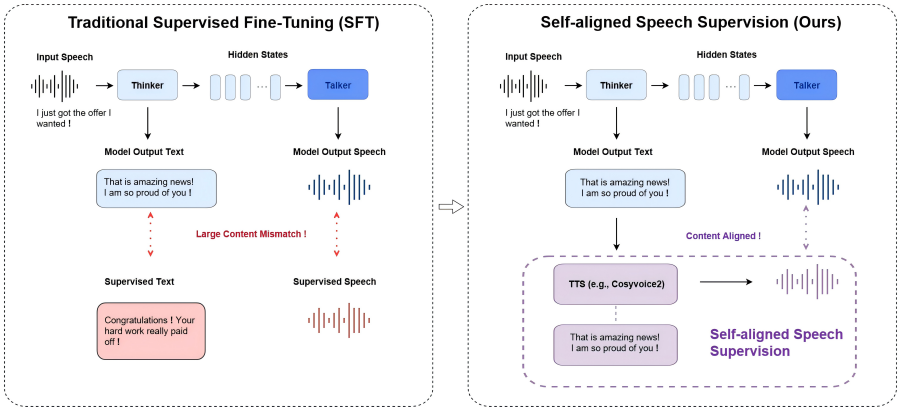

Fig. 3. Self-aligned supervision aligns speech targets with semantic representations produced by the Thinker module, enabling the Talker module to focus on prosody and pronunciation modeling.

By aligning speech supervision with the model's semantic predictions, the proposed strategy reduces semantic mismatch between hidden representations

and speech targets, allowing the Talker module to focus on speech generation rather than semantic correction. Figure 3 provides a comparison between conventional supervised fine-tuning and the proposed self-aligned speech supervision strategy.

4 Experiments

4.1 Datasets

Based on a high-quality open-source speech dialogue dataset [17], we construct dialect speech dialogue corpora for three Chinese dialects, including Sichuanese, Cantonese, and the Tianjin dialect, using the proposed synthesis pipeline. Each dialect contains approximately 3,600 speech query-response pairs. In addition, 3,600 high-MOS speech dialogue samples are selected for both Mandarin and English from this dataset and mixed with the constructed dialect data for joint training. The final training set contains approximately 18,000 speech dialogue pairs with a total duration of 225 hours, primarily consisting of open-domain daily conversations to better reflect real-world multilingual and dialect interaction scenarios.

4.2 Training Setup

We initialize the model from OpenS2S [17] due to its fully open-source nature and strong conversational capabilities in everyday spoken interactions. We first perform two epochs of supervised fine-tuning to obtain a base model with stable dialect capabilities, followed by self-aligned speech supervision training on the same dataset. All experiments are conducted on 8 NVIDIA A800 GPUs using DeepSpeed [12] with ZeRO Stage-2 [11]. The model is optimized using AdamW [10] with a learning rate of 2×10^{-5}, $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 1 \times 10^{-8}$, and weight decay 0.05. The learning rate schedule follows WarmupDecayLR.

5 Evaluation

5.1 Evaluation Setup

We evaluate DialectS2S on multilingual speech interaction from three aspects: language matching accuracy, response quality, and speech intelligibility. A multilingual dialect speech dialogue benchmark containing 250 test samples is constructed, covering two mainstream languages and three Chinese dialects, with evaluation queries built using different reference audios and query texts. Comparisons are conducted against several representative open-source speech dialogue models, including GLM-4-Voice [23], Qwen2.5-Omni [19], Kimi-Audio [6], Step-Audio2 [18], OpenS2S [17], and MiniCPM-o-4.5 [3].

5.2 Response Language Matching Accuracy

We first evaluate whether the model can generate speech responses in the input language or dialect while preserving multilingual capability. Specifically, we use FireRedLID [21] to identify the language of generated speech responses. All evaluations are conducted in a zero-shot setting without text prompts or explicit language control signals. Language matching accuracy is computed as the proportion of valid samples whose predicted language matches the input language. Results are shown in Table 1. DialectS2S achieves substantial improvements across all dialect scenarios while maintaining strong Mandarin and English performance, significantly outperforming existing open-source baselines.

Table 1. Comparison of language matching accuracy (%) between DialectS2S and baseline models.

Model	**Language Matching Accuracy**					
	English	Mandarin	Sichuanese	Cantonese	Tianjin	Avg.
GLM-4-Voice	94.00	97.92	2.22	46.94	0.00	48.22
Qwen2.5-Omni	100.00	100.00	0.00	0.00	0.00	40.00
Kimi-Audio	72.00	100.00	16.00	0.00	0.00	37.60
Step-Audio2	100.00	86.00	0.00	10.00	0.00	39.20
OpenS2S	100.00	93.88	0.00	0.00	0.00	38.78
MiniCPM-o-4.5	100.00	98.00	2.04	52.21	2.00	50.85
DialectS2S	100.00	97.96	**96.00**	**95.00**	**91.00**	**95.99**

5.3 Response Quality

Response quality is further evaluated on the constructed multilingual dialect benchmark using the text responses synchronously generated during speech interaction. Qwen3-Plus [22] is employed as the automatic evaluator to assess naturalness, semantic accuracy, and conformity to dialectal language usage, with scores ranging from 1 to 5. The scoring rubric is provided in Appendix A. Since all compared models support simultaneous text and speech generation, no additional ASR system is required. As shown in Table 2, DialectS2S achieves the highest overall average score and consistently delivers superior response quality across dialect scenarios, demonstrating strong semantic understanding and dialect-aware response generation capability.

Table 2. Response quality evaluation across different language and dialect scenarios.

Model	Response Quality					
	English	Mandarin	Sichuanese	Cantonese	Tianjin	Avg.
GLM-4-Voice	4.56	4.40	3.10	4.17	2.19	3.68
Qwen2.5-Omni	4.60	4.65	3.12	3.26	3.11	3.75
Kimi-Audio	4.07	4.12	3.22	3.80	2.26	3.49
Step-Audio2	4.20	3.92	2.78	3.02	2.06	3.20
OpenS2S	4.54	4.36	2.96	3.26	2.60	3.54
MiniCPM-o-4.5	4.22	4.45	3.51	4.24	2.30	3.74
DialectS2S	4.60	4.33	**4.48**	**4.28**	**4.40**	**4.42**

5.4 Speech Intelligibility

We use character error rate (CER) to measure speech intelligibility, where lower CER indicates more intelligible speech outputs. The text responses generated by the model are used as references, while FRASR2-AED [21] transcribes the generated speech responses for CER computation. Previous experiments show that existing baseline models mainly generate Mandarin Chinese rather than dialect speech under dialect interaction settings. Therefore, CER evaluation is instead conducted against the average CER reported by the FRASR2-AED dialect benchmark.

As shown in Table 3, DialectS2S achieves low CER across both Mandarin and dialect scenarios, demonstrating strong speech intelligibility. In particular, DialectS2S achieves lower CER than the average performance reported by the FRASR2-AED dialect benchmark in dialect settings, indicating highly intelligible generated dialect speech.

Table 3. CER (%) of generated speech compared with the average CER reported by the FRASR2-AED benchmark.

	Mandarin	Dialect
DialectS2S	3.28	**8.69**
Benchmark Avg.	3.05	11.67

6 Ablation Study

To evaluate the effectiveness of the proposed self-aligned speech supervision strategy, we conduct ablation studies under three training configurations: **sft-2ep**, which applies two epochs of supervised fine-tuning; **sft-3ep**, which continues supervised fine-tuning for one additional epoch; and **DialectS2S**, which applies self-aligned speech supervision on top of the sft-2ep model. Notably, sft-3ep and DialectS2S are trained with comparable numbers of optimization steps, enabling a fair comparison between continued conventional fine-tuning and the proposed training strategy.

We evaluate both response quality and speech intelligibility, with results reported in Table 4 and Table 5, respectively. As shown in Table 4, DialectS2S achieves response quality comparable to sft-3ep, indicating that self-aligned

speech supervision does not compromise semantic generation capability. In contrast, Table 5 shows that DialectS2S consistently outperforms both sft-2ep and sft-3ep in speech intelligibility. In particular, simply extending supervised fine-tuning does not yield consistent improvements in CER and may even lead to degradation in certain dialect settings.

These results suggest that increasing the amount of supervised fine-tuning alone is insufficient to address the hidden-state distribution shift in the Thinker-Talker architecture. In comparison, the proposed self-aligned speech supervision explicitly aligns speech targets with the model's semantic predictions, thereby reducing semantic mismatch and improving speech clarity and stability.

Table 4. Response quality evaluation under different training strategies across different language and dialect scenarios.

Model	**Response Quality**					
	English	Mandarin	Sichuanese	Cantonese	Tianjin	Avg.
sft-2ep	3.97	4.19	4.02	3.81	4.36	4.07
sft-3ep	4.27	4.40	4.67	4.19	4.57	**4.42**
DialectS2S	4.60	4.33	4.48	4.28	4.40	**4.42**

Table 5. Comparison of CER (%) under different training strategies.

Model	**ASR CER ↓**					
	English	Mandarin	Sichuanese	Cantonese	Tianjin	Avg.
sft-2ep	8.68	5.25	11.29	17.10	8.43	10.15
sft-3ep	11.49	6.85	9.85	19.80	8.36	11.27
DialectS2S	5.56	3.28	8.63	9.53	7.92	**6.98**

7 Conclusion

This work presents DialectS2S, an end-to-end speech dialogue model for low-resource Chinese dialects. By combining a scalable dialect data synthesis pipeline with a two-stage post-training strategy, DialectS2S enables stable and intelligible dialect speech interaction under limited supervision. Experimental results demonstrate strong language and dialect matching, response quality, and text–speech consistency across three Chinese dialects. To facilitate future research and practical applications, we fully open-source the DialectS2S framework, including model checkpoints, training datasets, and fine-tuning code. In future work, we plan to support more dialects and explore more effective training strategies for improving dialect authenticity and expressiveness.

References

1. An, K., Chen, Y., Chen, Z., Deng, C., Du, Z., Gao, C., Gao, Z., Gong, B., Li, X., Li, Y., et al.: Fun-asr technical report. arXiv preprint arXiv:2509.12508 (2025)
2. Chen, Z., Chen, G., Wang, Y., Ding, C., Zhang, W.Q., et al.: Diamoe-tts: A unified ipa-based dialect tts framework with mixture-of-experts and parameter-efficient zero-shot adaptation. arXiv preprint arXiv:2509.22727 (2025)
3. Cui, J., Xu, B., Wang, C., Yu, T., Sun, W., Xu, Y., Wang, T., He, Z., Ma, W., Cai, T., et al.: Minicpm-o 4.5: Towards real-time full-duplex omni-modal interaction. arXiv preprint arXiv:2604.27393 (2026)
4. Dai, Y., Zhang, Z., Wang, S., Li, L., Guo, Z., Zuo, T., Wang, S., Xue, H., Wang, C., Wang, Q., Xu, X., Bu, H., Li, J., Kang, J., Zhang, B., Xie, L.: Wenetspeech-chuan: A large-scale sichuanese corpus with rich annotation for dialectal speech processing. In: ICASSP 2026 - 2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 19507–19511 (2026). https://doi.org/10.1109/ICASSP55912.2026.11463960
5. Défossez, A., Mazaré, L., Orsini, M., Royer, A., Pérez, P., Jégou, H., Grave, E., Zeghidour, N.: Moshi: a speech-text foundation model for real-time dialogue. arXiv preprint arXiv:2410.00037 (2024)
6. Ding, D., Ju, Z., Leng, Y., Liu, S., Liu, T., Shang, Z., Shen, K., Song, W., Tan, X., Tang, H., et al.: Kimi-audio technical report. arXiv preprint arXiv:2504.18425 (2025)
7. Du, Z., Wang, Y., Chen, Q., Shi, X., Lv, X., Zhao, T., Gao, Z., Yang, Y., Gao, C., Wang, H., et al.: Cosyvoice 2: Scalable streaming speech synthesis with large language models. arXiv preprint arXiv:2412.10117 (2024)
8. Fang, Q., Guo, S., Zhou, Y., Ma, Z., Zhang, S., Feng, Y.: Llama-omni: Seamless speech interaction with large language models. In: International Conference on Learning Representations. vol. 2025, pp. 57607–57624 (2025)
9. Li, L., Guo, Z., Chen, H., Dai, Y., Zhang, Z., Xue, H., Zuo, T., Wang, C., Wang, S., Xu, X., Bu, H., Li, J., Kang, J., Zhang, B., Yuan, R., Zhou, Z., Xue, W., Xie, L.: Wenetspeech-yue: A large-scale cantonese speech corpus with multi-dimensional annotation. Proceedings of the AAAI Conference on Artificial Intelligence **40**, 31627–31635 (Mar 2026). https://doi.org/10.1609/aaai.v40i37.40429
10. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101 (2017)
11. Rajbhandari, S., Rasley, J., Ruwase, O., He, Y.: Zero: Memory optimizations toward training trillion parameter models. In: SC20: international conference for high performance computing, networking, storage and analysis. pp. 1–16. IEEE (2020)
12. Rasley, J., Rajbhandari, S., Ruwase, O., He, Y.: Deepspeed: System optimizations enable training deep learning models with over 100 billion parameters. In: Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining. pp. 3505–3506 (2020)
13. Saeki, T., Xin, D., Nakata, W., Koriyama, T., Takamichi, S., Saruwatari, H.: UTMOS: UTokyo-SaruLab System for VoiceMOS Challenge 2022. In: Interspeech 2022. pp. 4521–4525 (2022). https://doi.org/10.21437/Interspeech.2022-439
14. Shi, Y., Shu, Y., Dong, S., Liu, G., Sesay, J., Li, J., Hu, Z.: Voila: Voice-language foundation models for real-time autonomous interaction and voice role-play. arXiv preprint arXiv:2505.02707 (2025)
15. Tang, Z., Wang, D., Xu, Y., Sun, J., Lei, X., Zhao, S., Wen, C., Tan, X., Xie, C., Zhou, S., et al.: Kespeech: An open source speech dataset of mandarin and its eight

subdialects. In: Thirty-fifth conference on neural information processing systems datasets and benchmarks track (Round 2) (2021)
16. Tian, F., Zhang, X.T., Zhang, Y., Zhang, H., Li, Y., Liu, D., Deng, Y., Wu, D., Chen, J., Zhao, L., et al.: Step-audio-r1 technical report. arXiv preprint arXiv:2511.15848 (2025)
17. Wang, C., Peng, T., Yang, W., Bai, Y., Wang, G., Lin, J., Jia, L., Wu, L., Wang, J., Zong, C., et al.: Opens2s: Advancing fully open-source end-to-end empathetic large speech language model. In: Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: System Demonstrations. pp. 906–917 (2025)
18. Wu, B., Yan, C., Hu, C., Yi, C., Feng, C., Tian, F., Shen, F., Yu, G., Zhang, H., Li, J., et al.: Step-audio 2 technical report. arXiv preprint arXiv:2507.16632 (2025)
19. Xu, J., Guo, Z., He, J., Hu, H., He, T., Bai, S., Chen, K., Wang, J., Fan, Y., Dang, K., et al.: Qwen2.5-omni technical report. arXiv preprint arXiv:2503.20215 (2025)
20. Xu, J., Guo, Z., Hu, H., Chu, Y., Wang, X., He, J., Wang, Y., Shi, X., He, T., Zhu, X., et al.: Qwen3-omni technical report. arXiv preprint arXiv:2509.17765 (2025)
21. Xu, K., Jia, Y., Huang, K., Chen, J., Li, W., Liu, K., Xie, F.L., Tang, X., Hu, Y.: Fireredasr2s: A state-of-the-art industrial-grade all-in-one automatic speech recognition system. arXiv preprint arXiv:2603.10420 (2026)
22. Yang, A., Li, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Gao, C., Huang, C., Lv, C., et al.: Qwen3 technical report. arXiv preprint arXiv:2505.09388 (2025)
23. Zeng, A., Du, Z., Liu, M., Wang, K., Jiang, S., Zhao, L., Dong, Y., Tang, J.: Glm-4-voice: Towards intelligent and human-like end-to-end spoken chatbot. arXiv preprint arXiv:2412.02612 (2024). https://doi.org/10.48550/arXiv.2412.02612
24. Zhang, Y., Du, Y., Dai, Z., Ma, X., Kou, K., Wang, B., Li, H.: Echox: Towards mitigating acoustic-semantic gap via echo training for speech-to-speech llms. arXiv preprint arXiv:2509.09174 (2025)

A Additional Experimental Details

Response-quality rubric. Qwen3-Plus evaluates every model's synchronously generated text with the same prompt, without an additional ASR step. The five-point scoring rubric is:

1: The response is completely irrelevant or severely fails to satisfy the user's request.
2: The response is partially relevant but contains clear errors or missing content that make it difficult to understand.
3: The response is generally relevant but contains some errors or unnatural expressions and does not fully satisfy the user's request.
4: The response is of high overall quality, with only minor issues, and is generally natural and easy to understand.
5: The response fully satisfies the user's request, is natural and fluent, is accurately expressed, and is highly consistent with the dialogue context.

For dialect inputs, the evaluator additionally considers whether the textual dialect usage is natural and consistent with the requested dialect.

Comparison with a cascaded system. DialectS2S outperforms a cascaded system with a comparable parameter count in both response quality and inference speed. The cascade baseline combines FireRedASR2, Qwen3-8B, and CosyVoice2. Response quality is evaluated by Qwen3-Plus using the five-point rubric described above.

Table 6. Supplementary cascade comparison and streaming-inference latency. Latency is measured on the first 50 Sichuanese samples using one NVIDIA A800-SXM4-80GB GPU.

(a) Cascade comparison

Metric	Cascade	DialectS2S
Success	49/50	**50/50**
Text startup (s)	1.170	**0.852**
Audio startup (s)	8.589	**5.110**
Quality (1–5)	2.15	**4.48**

(b) Streaming latency (s)

Metric	Mean	Median	P90	P95
TTFT	0.852	0.532	0.671	3.698
First audio	5.110	4.879	5.399	8.759
Full generation	42.976	44.225	76.493	80.753