SoulX-Podcast: Towards Realistic Long-form Podcasts with Dialectal and Paralinguistic Diversity

Hanke Xie[1,2*†] **Haopeng Lin**[2†] **Wenxiao Cao**[2] **Dake Guo**[1] **Wenjie Tian**[1]
Jun Wu[2] **Hanlin Wen**[2] **Ruixuan Shang**[2] **Hongmei Liu**[2] **Zhiqi Jiang**[2]
Yuepeng Jiang[1] **Wenxi Chen**[2,3*] **Ruiqi Yan**[2,3*] **Jiale Qian**[2] **Yichao Yan**[2]
Shunshun Yin[2] **Ming Tao**[2] **Xie Chen**[3] **Lei Xie**[1‡] **Xinsheng Wang**[2‡]

[1] Audio, Speech and Language Processing Group (ASLP@NPU),
Northwestern Polytechnical University, Xi'an, China
[2] Soul AI Lab, China
[3] X-LANCE Lab, Shanghai Jiao Tong University, China

Abstract

Recent advances in text-to-speech (TTS) synthesis have significantly improved speech expressiveness and naturalness. However, most existing systems are tailored for single-speaker synthesis and fall short in generating coherent multi-speaker conversational speech. This technical report presents SoulX-Podcast, a system designed for podcast-style multi-turn, multi-speaker dialogic speech generation, while also achieving state-of-the-art performance in conventional text-to-speech (TTS) tasks. To meet the higher naturalness demands of multi-turn spoken dialogue, SoulX-Podcast integrates a range of paralinguistic controls and supports both Mandarin and English, as well as several Chinese dialects, including Sichuanese, Henanese, and Cantonese, enabling more personalized podcast-style speech generation. Experimental results demonstrate that SoulX-Podcast can continuously produce over 90 minutes of conversation with stable speaker timbre and smooth speaker transitions. Moreover, speakers exhibit contextually adaptive prosody, reflecting natural rhythm and intonation changes as dialogues progress. Across multiple evaluation metrics, SoulX-Podcast achieves state-of-the-art performance in both monologue TTS and multi-turn conversational speech synthesis.

Demo page: https://soul-ailab.github.io/soulx-podcast
Source code: https://github.com/Soul-AILab/SoulX-Podcast

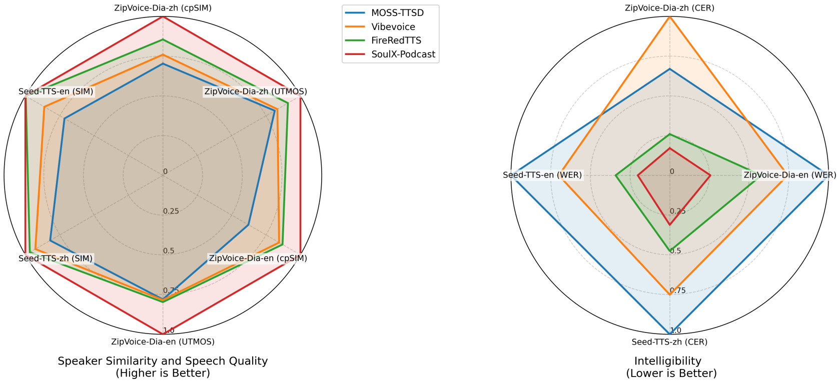

Figure 1: Performance of SoulX-Podcast.

[*] Work done during an internship at Soul AI Lab. `hkxie@mail.nwpu.edu.cn`

[†] Equal contribution.

[‡] Corresponding authors. `lxie@nwpu.edu.cn`, `wangxinsheng@soulapp.cn`

Technical Report.

1 Introduction

Building upon the generative power of large language models (LLMs), modern text-to-speech (TTS) systems have reached a point where they can produce speech that is nearly indistinguishable from human voices, achieving remarkable naturalness and zero-shot voice cloning performance [1], [2], [3], [4], [5]. However, most previous work has primarily focused on single-speaker speech generation. These systems, while effective in isolated speech tasks, struggle to maintain fluency and naturalness in multi-speaker, multi-turn conversation scenarios. In response to this gap, this technical report introduces **SoulX-Podcast**, a speech synthesis model specifically designed for seamless multi-speaker, multi-turn dialogues. To further enhance conversational realism and diversity, SoulX-Podcast incorporates robust support for various paralinguistic features and dialects, ensuring a more dynamic and natural dialogue experience.

Speech tokenization methods [6], [7], [8], [9], [10], [11], [12], [13], [14], based on Vector Quantization (VQ) [15] or Finite Scalar Quantization [16], bridge the gap between continuous speech signals and discrete token-based large language models (LLMs). Vall-E [17] is a pioneering speech generation system that leverages LLMs with discrete tokens and employs a tokenizer based on residual vector quantization (RVQ) [18]. Specifically, tokens from the first layer are predicted by an autoregressive (AR) LLM, while the remaining tokens are generated using a non-autoregressive (NAR) model. Subsequent work, depending on the choice of tokenizer, can be categorized into several main modeling approaches: predicting a single stream of semantic tokens with LLMs and then generating acoustic features via flow matching [5], [19], [20], [21], [22], [23]; directly predicting acoustic tokens spanning multiple codebooks according to specific patterns [24], [25]; or directly predicting a single stream of acoustic tokens [2], [3].

Most of the aforementioned research primarily focuses on monologue-style speech generation, overlooking the challenges of multi-speaker, multi-turn conversational synthesis. In contrast, conversational speech generation places higher demands on natural prosodic and rhythmic variation to ensure smooth and coherent dialogue flow. Recently, several studies have begun to explore this direction. For example, Covomix [26] adopts a parallel-channel modeling strategy that simultaneously predicts the speech of different speakers through separate channels, while MoonCast [27] and MOSS-TTSD [28] merge dialogue text with speaker labels to generate integrated multi-speaker conversations. The latest FireRedTTS-2 [25] model, on the other hand, produces speech from multiple speakers in an alternating manner. Although these methods achieve improved dialogue continuity and prosodic variation compared with standard TTS systems, their limited control over paralinguistic features still constrains the expressiveness and realism of the generated conversations.

In this work, we introduce SoulX-Podcast, a large language model–driven framework for long-form, multi-speaker, and multi-dialect podcast speech synthesis. SoulX-Podcast is designed to generate stable, coherent, and expressive podcast-style dialogic speech by effectively modeling dialectal variation, paralinguistic cues, and context-dependent prosody. The framework represents interleaved text–speech sequences, where speaker-labeled text and corresponding speech tokens are chronologically aligned, thereby facilitating the generation of long-form conversational audio with consistent quality and speaker similarity. Experimental results demonstrate that SoulX-Podcast delivers superior performance in multi-turn dialogue synthesis and exhibits strong generalization to conventional TTS tasks, highlighting its versatility across diverse speech generation scenarios.

In summary, SoulX-Podcast offers several distinctive features:

- It supports **long-form, natural dialogue speech generation** with a variety of **paralinguistic labels**, achieving high fluency and coherence across extended multi-turn dialogues.
- In addition to **Mandarin** and **English**, SoulX-Podcast provides robust support for several Chinese dialects, including **Sichuanese**, **Henanese**, and **Cantonese**, enabling more diverse and personalized voice generation. Importantly, all of these dialects support **Cross-dialectal, zero-shot voice cloning**, allowing a single audio prompt to generate speech in any of the supported dialects.
- SoulX-Podcast demonstrates superior performance not only in multi-turn conversational speech synthesis but also in conventional TTS tasks, such as **voice cloning**, highlighting its effectiveness and versatility across diverse speech synthesis scenarios.

2 Method

Dialogue text–speech paired data that contain speaker identity correspondence are a necessary prerequisite for building a dialogue speech synthesis system. This section first introduces the data processing methods used in this work, including the handling of dialogue data and the annotation of dialectal and paralinguistic information. Subsequently, we present the specific algorithm of SoulX-Podcast.

2.1 Data Processing

In contrast to monologue speech, the processing of dialogue speech necessitates not only obtaining aligned transcripts but also distinguishing between speakers explicitly. As shown in Figure 2, the overall workflow comprises speech enhancement, audio segmentation and speaker diarization, text transcription, and quality filtering. Additionally, to facilitate paralinguistic and dialectal controllability, further information is extracted and annotated.

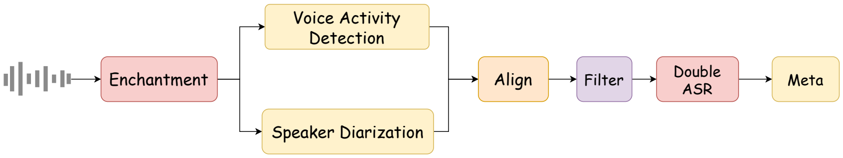

Figure 2: Processing pipeline for in-the-wild dialogue speech data.

2.1.1 Basic Processing Workflow

Audio Pre-processing. In-the-wild dialogue recordings often contain background music or noise, which can adversely affect downstream tasks such as transcription or speaker diarization. To address this, we first apply a UVR-MDX-based[4] vocal separation tool to remove background audio and noise, and then normalize the resulting signals to a consistent amplitude.

Segmentation and Diarization. Processing long-form dialogue recordings (e.g., longer than 30 minutes) poses challenges for conventional speaker diarization. As speakers' vocal characteristics and speaking states vary over time, diarization models may mistakenly assign multiple speaker identities to the same individual, leading to inconsistencies in speaker counts and boundary alignment.

To mitigate this issue, we first apply Voice Activity Detection (VAD) [29] to segment long recordings into short utterances. These utterances are then concatenated into dialogue segments of approximately five minutes. During this process, we enforce a silence-duration constraint to prevent segment boundaries from crossing different sessions or long transitional silences: if the inter-utterance silence exceeds a predefined threshold, the adjacent utterances are treated as the end and start of separate segments.

Finally, we employ a `Sortformer`-based diarization model [30] to detect speaker boundaries and assign speaker labels, producing reliable speaker-turn annotations for subsequent processing.

Quality Filtering. Although the audio recordings were enhanced in the initial stage, some segments still exhibited suboptimal denoising results or inherently poor recording quality. To prevent such low-quality data from negatively impacting model training, we applied a series of filtering criteria to the dialogue segments, including signal-to-noise ratio (SNR) and perceptual quality estimated by DNSMOS [31].

Speech Recognition. Following the quality filtering process, we employed a dual-ASR transcription strategy to obtain reliable transcripts. Specifically, each utterance within a dialogue segment was transcribed by two independent ASR models. For Chinese speech, we used `Paraformer`[5] and `Whisper`[6], while for English speech, we adopted `Parakeet`[7] and `Whisper`.

[4] `https://github.com/seanghay/uvr-mdx-infer`

[5] https://huggingface.co/funasr/Paraformer-large

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

[7] https://huggingface.co/nvidia/parakeet-tdt-0.6b-v2

For each utterance, two transcription results were obtained, and the Character Error Rate (CER) for Chinese or Word Error Rate (WER) for English was computed. Utterances with CER or WER below a predefined threshold were fully retained, with Paraformer outputs used as the final transcripts for Chinese and Whisper outputs used for English. For utterances whose CER or WER exceeded the threshold, only the textual transcripts were preserved, while the corresponding audio was discarded.

This strategy maintains dialogue completeness and textual consistency while minimizing the adverse impact of transcription errors on speech synthesis training, thereby achieving a better trade-off between data retention and transcription reliability.

Speaker Purity Refinement. To ensure speaker label consistency, we conducted a speaker-purity refinement based on speaker embedding clustering. For each dialogue segment, the embeddings of all utterances belonging to the same speaker were clustered, and utterances whose embeddings deviated excessively from the cluster centroid were identified as outliers. These outlier utterances were excluded from the audio data—only their transcriptions were retained. This strategy effectively mitigates potential speaker confusion during multi-turn dialogue synthesis while maximizing overall data retention. Here, we extract speaker embeddings with WavLM-large, finetuned on the speaker verification task [32].

2.1.2 Paralinguistic and Dialectal Data Annotation

Paralinguistic cues, such as laughter and sighs, play a crucial role in enhancing the naturalness and expressiveness of dialogue. To enable controllable generation of such cues, we performed paralinguistic mining and annotation on the collected data. Moreover, most previous speech synthesis research has primarily focused on Mandarin Chinese, while major Chinese dialects such as Cantonese and Sichuanese have received limited attention. To facilitate effective dialectal controllability, we further annotated the collected data with dialectal labels, enabling the model to capture and reproduce dialect-specific characteristics.

Paralinguistic Annotation. To ensure both large-scale coverage and fine-grained precision of paralinguistic labels, we design a two-stage refinement framework for data annotation. This framework combines high-throughput automated detection with model-assisted verification to achieve both efficiency and accuracy.

In the first stage, we employ language-specific ASR models fine-tuned for paralinguistic event detection to process the raw audio corpus. For Mandarin Chinese data, we use `Beats` [33] for coarse identification of nonverbal cues, while for English data, we adopt `Whisperd` [34]. This stage efficiently filters out segments unlikely to contain relevant paralinguistic events.

In the second stage, the pre-annotated segments undergo model-driven verification and fine-grained labeling using the Gemini-2.5-Pro API [35]. This multimodal model verifies the presence and category of paralinguistic events and generates precise, time-aligned annotations alongside the corresponding text.

Through this meticulous two-stage process, we obtain approximately 1,000 hours of high-quality speech with detailed paralinguistic annotations, providing a robust foundation for expressive and context-aware speech synthesis.

Dialectal Annotation. To efficiently collect dialectal speech data, we employed two complementary strategies. First, we collected publicly available recordings in specific Chinese dialects. Second, we trained a dialect identification model to retrieve and categorize dialectal utterances from the broader in-the-wild dataset.

For transcription, we observed that our standard pipeline performed suboptimally on dialectal speech. Accordingly, we leveraged the commercial Seed-ASR API [8] to generate reliable transcripts. Using this approach, we obtained approximately 2,000 hours of Sichuanese, 1,000 hours of Cantonese, and 500 hours of Henanese speech.

2.1.3 Corpus Overview

Using the aforementioned methods, we ultimately obtained approximately **0.3 million hours** of high-quality, natural conversational speech. In addition, we curated roughly 1.0 million hours of monologue data, resulting in a total training dataset of approximately **1.3 million hours**.

[8] https://docs.byteplus.com/zh-CN/docs/byteplusvoice/asrstreaming

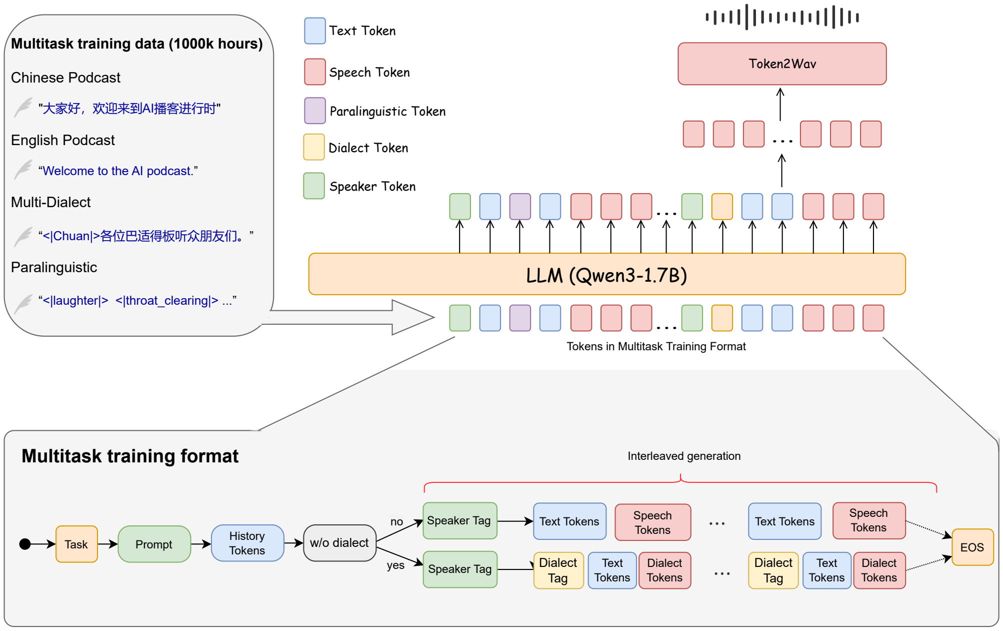

Figure 3: Overview of SoulX-Podcast

2.2 SoulX-Podcast

Following the CosyVoice series [5], [23], SoulX-Podcast adopts a two-stage generative framework. Specifically, an LLM first predicts semantic tokens, which are then converted into acoustic features through flow matching and subsequently synthesized into waveform audio via a vocoder. The LLM backbone is the pre-trained Qwen3-1.7B model, whose text codebook is extended to accommodate both speech tokens and special tokens that encode paralinguistic and dialectal attributes.

2.2.1 Token Organization

To enable flexible, multi-turn dialogue generation, we adopt a text–speech interleaved sequence that allows sentence-by-sentence synthesis. Specifically, each speaker's text tokens are followed by their corresponding speech tokens, which are then concatenated with the next speaker's text and speech tokens in temporal order. Each utterance begins with a speaker token to indicate the speaker identity. Likewise, dialect control is achieved by inserting a dialect-specific token immediately after the speaker token, while paralinguistic cues (e.g., laughter, sighs) are treated as textual tokens and placed at their corresponding positions within the sequence. An example with a dialect label is shown below:

```
<SPEAKER1><Sichuan><Text Tokens><Audio Tokens><SPEAKER2><Sichuan><Text
Tokens><Audio Tokens><SPEAKER3><...>
```

2.2.2 Training

Dialogue speech data are relatively scarce compared to monologue speech. To effectively leverage heterogeneous data patterns and enhance performance in dialogue scenarios, we adopt a curriculum learning strategy.

In the first stage, the LLM backbone is initialized from Qwen3-1.7B [9] and trained on a mixture of monologue and dialogue data to acquire fundamental text-to-speech capabilities. Subsequently, the model is further trained on multi-speaker dialogue data in both Chinese and English, incorporating dialectal and paralinguistic elements. Since the amount of Chinese dialect data is significantly smaller

[9] https://huggingface.co/Qwen/Qwen3-1.7B

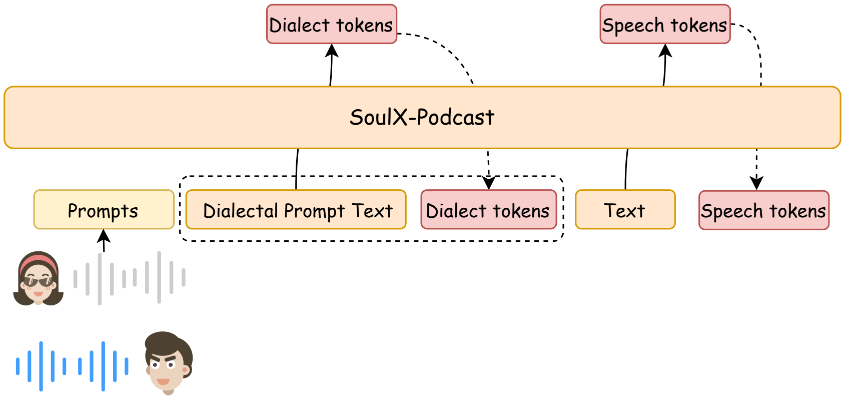

Figure 4: Inference procedure of SoulX-Podcast. The model supports Cross-dialectal prompting, where a Mandarin prompt can generate speech in target dialects with Dialect-Guided Prompting (DGP) method.

than that of Mandarin and English, we perform additional fine-tuning on dialectal data to enhance the model's dialectal capability, resulting in a podcast model specifically optimized for dialect generation.

To address the challenges of long-form audio generation, we introduce a context regularization mechanism that progressively drops historical speech tokens while retaining their textual context. This encourages the model to rely on semantic continuity rather than low-level acoustic memory, thereby improving coherence and stability in extended conversational synthesis.

2.2.3 Inference

During inference, we follow the token organization established in training: initial text and speech tokens from multiple speakers are interleaved, and the model autoregressively generates subsequent speech tokens in the same interleaved manner.

Cross-dialectal Voice Cloning. For dialect generation, our goal is to enable cross-dialectal voice cloning. However, this is nontrivial. Unlike the clear orthographic differences between Chinese and English, various Chinese dialects—particularly Mandarin, Henanese, and Sichuanese—share an identical written form. Even Cantonese, though linguistically more distinct, still exhibits substantial textual overlap with Mandarin. Consequently, when the target text is highly similar to Mandarin and the speech prompt is also in Mandarin, the dialectal control signal becomes weak.

To address this issue and allow a Mandarin prompt to generate speech in any target dialect, we propose **Dialect-Guided Prompting (DGP)** inference strategy. Specifically, before generating a dialectal podcast, we prepend a short dialect-typical sentence—one that strongly reflects the target dialectal style—to the input text. This initial utterance effectively guides the model toward producing speech with the desired dialectal characteristics in subsequent generations.

3 Performance of SoulX-Podcast

Although SoulX-Podcast is designed for multi-turn, multi-speaker dialogue synthesis, it is also capable of conventional monologue speech synthesis. Accordingly, we first compare its performance against SOTA TTS models on the standard monologue synthesis task. We then evaluate SoulX-Podcast's capabilities in dialogue generation, as well as in paralinguistic and dialectal control.

3.1 Monologue Speech Generation

To evaluate the zero-shot (voice cloning) TTS capability of SoulX-Podcast, we assess its performance on Seed-TTS-eval and compare it with existing zero-shot TTS models. The results are summarized in Table 1, where speech intelligibility is measured using CER for Chinese and WER for English, and speaker similarity (SIM) is quantified via the cosine similarity of speaker embeddings, following the Seed-TTS-eval protocol[10]. As can been seen, SoulX-Podcast demonstrates significant superiority in intelligibility for zero-shot monologue TTS scenarios. Specifically, SoulX-Podcast achieves lowest CER in the Chinese test set. In the English test set, SoulX-Podcast only seconds to F5-TTS. In terms of speaker similarity, SoulX-Podcast also achieves strong results. Specifically, on both the Chinese and English test sets, it ranks just behind Seed-TTS and MaskGCT, demonstrating its excellent performance in conventional zero-shot TTS.

Table 1: TTS performance of different models on the Seed test sets (test-zh for Chinese, test-en for English). Arrows indicate the desired direction (↓ = lower is better, ↑ = higher is better). Best values per column are in **bold**.

Model	**test-zh**		**text-en**	
	CER (↓)	SIM (↑)	WER (↓)	SIM (↑)
Seed-TTS [21]	1.12	**0.796**	2.25	**0.762**
MaskGCT [24]	2.27	0.774	2.62	0.714
F5-TTS [36]	1.56	0.741	**1.83**	0.647
CosyVoice2 [23]	1.45	0.748	2.57	0.652
Llasa-8B-250k [2]	1.59	0.684	2.97	0.574
Spark-TTS [3]	1.20	0.672	1.98	0.584
MOSS-TTSD [28]	3.53	0.609	9.47	0.473
Vibevoice [37]	2.65	0.689	6.62	0.570
FireRedTTS2 [25]	1.68	0.719	3.23	0.659
SoulX-Podcast	**1.10**	0.743	1.91	0.661

3.2 Podcast Generation

To evaluate multi-turn, multi-speaker dialogue generation, we compare SoulX-Podcast with representative dialogue TTS systems on the ZipVoice-Dia test set. This benchmark comprises natural multi-turn conversations, enabling assessment of both intelligibility and cross-speaker consistency (cpSIM) in long-form synthesis. As shown in Table 2, SoulX-Podcast outperforms recent state-of-the-art models on both the Chinese and English subsets. Specifically, it achieves the lowest WER/CER and the highest cpSIM, while maintaining competitive UTMOS scores, demonstrating superior speaker coherence and perceived quality.

Table 2: Objective evaluation of multi-speaker TTS systems on ZipVoice-Dia test sets. Arrows indicate the desired direction (↓ = lower is better, ↑ = higher is better). Best values per column are in **bold**.

Model	**ZipVoice-Dia (zh)**			**ZipVoice-Dia (en)**		
	CER (↓)	cpSIM (↑)	UTMOS (↑)	WER (↓)	cpSIM (↑)	UTMOS (↑)
ZipVoice-Dia [38]	3.39	0.553	**2.24**	3.32	0.438	**3.10**
MoonCast [27]	27.43	0.441	1.76	23.62	0.356	2.30
MOSS-TTSD [28]	8.62	0.421	1.70	8.86	0.301	2.31
Vibevoice [37]	12.87	0.455	1.74	6.58	0.409	2.32
FireRedTTS2 [25]	3.34	0.512	1.90	5.11	0.421	2.36
SoulX-Podcast	**2.2**	**0.599**	2.09	**2.27**	**0.484**	2.96

3.3 Evaluation of Paralinguistic Control

To evaluate the proposed model's capability for controllable paralinguistic generation, we constructed a dedicated paralinguistic test set. Concretely, we employed GPT-5 to generate 20 test utterances for each of five paralinguistic labels, i.e., `<|laughter|>`, `<|sigh|>`, `<|breathing|>`, `<|coughing|>`,

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

and <throat_clearing>. The corresponding audio samples were subsequently synthesized using SoulX-Podcast in monologue speech synthesis mode.

For objective evaluation, we adopted the Qwen-2.5 Omni-FT model [39] as an automated paralinguistic recognizer. The evaluator was tasked with verifying whether each synthesized utterance contained the target paralinguistic event specified in the prompt. The resulting recognition accuracies are summarized in Table 3.

Table 3: Recognition accuracy for different paralinguistic labels.

Label	Count	Correct	Error	Accuracy
laughter	20	20	0	1.00
sigh	20	17	3	0.85
breathing	20	15	5	0.75
coughing	20	14	6	0.70
throat_clearing	20	16	4	0.80
Total / Average	100	82	18	0.82

As shown in Table 3, our model achieves a strong overall accuracy of 0.82 in controlling these paralinguistic events. It demonstrates near-perfect control over distinct events like <|laughter|> and high fidelity for <|sigh|> and <|throat_clearing|>. The primary sources of error appear concentrated in more acoustically subtle or ambiguous events, namely <|breathing|> (0.75) and <|coughing|> (0.70), which may be more challenging for the evaluator model to distinguish.

3.4 Dialect Generation

SoulX-Podcast currently supports three major Chinese dialects: Sichuanese, Henanese, and Cantonese. We evaluate its performance on these dialects in both monologue TTS and dialogue generation settings. The monologue test set includes 1,000 samples per dialect, drawn from internal GPT-generated data as well as SeedTTS, Wenetspeech-Yue-eval [40], and Wenetspeech-Chuan-eval [41]. The dialogue test set contains 100 GPT-generated items per dialect.

Dialect-specific ASR systems are used to compute CER, including Wenetspeech-Chuan-ASR [41] for Sichuanese, TeleSpeech[11] for Henanese, and Wenetspeech-Yue-ASR [40] for Cantonese. As shown in Table 4, SoulX-Podcast achieves consistent speaker similarity across all three dialects, comparable to its performance on Mandarin and English. The relatively high CER values may partly arise from limitations of the ASR systems.

Table 4: Performance evaluation of SoulX-Podcast on TTS and dialogue generation across different dialects. Arrows indicate the desired direction (↓ = lower is better, ↑ = higher is better).

Dialect	**Monologue Test**		**Dialogue Test**	
	CER (↓)	SIM (↑)	CER (↓)	cpSIM (↑)
Sichuanese	3.75	0.704	15.42	0.641
Henanese	8.14	0.705	28.06	0.647
Cantonese	9.77	0.680	19.50	0.627

4 Conclusions

In this work, we introduced SoulX-Podcast, a large language model–driven framework for long-form, multi-speaker, and multi-dialect conversational speech synthesis. Through an interleaved text–speech modeling paradigm, SoulX-Podcast enables the generation of long-form, multi-turn conversational speech with consistent quality and coherence. Experimental results demonstrate that SoulX-Podcast not only excels in multi-turn dialogue synthesis but also generalizes effectively to zero-shot monologue TTS. Its capability to handle multiple Chinese dialects and paralinguistic cues further highlights its versatility and potential as a unified framework for speech generation.

[11] https://github.com/Tele-AI/TeleSpeech-ASR

5 Ethics Statement

This work focuses on advancing speech synthesis technology through the development of SoulX-Podcast, a large language model–driven framework for multi-speaker and multi-dialect conversational speech generation. All datasets used in this study were either publicly available or synthetically generated, and no personally identifiable information or private recordings were included.

We acknowledge the potential risks associated with misuse of speech synthesis technology, such as voice spoofing, impersonation, or misinformation. To mitigate these risks, SoulX-Podcast is intended solely for research and responsible development of speech interfaces, and any downstream applications should incorporate appropriate speaker consent, watermarking, and misuse detection mechanisms. We advocate for the transparent, ethical, and human-centric use of speech generation technologies.

References

[1] H.-H. Guo, Y. Hu, K. Liu, F.-Y. Shen, X. Tang, Y.-C. Wu, F.-L. Xie, K. Xie, and K.-T. Xu, "Fireredtts: A foundation text-to-speech framework for industry-level generative speech applications," *arXiv preprint arXiv:2409.03283*, 2024.

[2] Z. Ye, X. Zhu, C.-M. Chan, X. Wang, X. Tan, J. Lei, Y. Peng, H. Liu, Y. Jin, Z. Dai, et al., "Llasa: Scaling train-time and inference-time compute for llama-based speech synthesis," *arXiv preprint arXiv:2502.04128*, 2025.

[3] X. Wang, M. Jiang, Z. Ma, Z. Zhang, S. Liu, L. Li, Z. Liang, Q. Zheng, R. Wang, X. Feng, et al., "Spark-tts: An efficient llm-based text-to-speech model with single-stream decoupled speech tokens," *arXiv preprint arXiv:2503.01710*, 2025.

[4] W. Deng, S. Zhou, J. Shu, J. Wang, and L. Wang, "Indextts: An industrial-level controllable and efficient zero-shot text-to-speech system," *arXiv preprint arXiv:2502.05512*, 2025.

[5] Z. Du, C. Gao, Y. Wang, F. Yu, T. Zhao, H. Wang, X. Lv, H. Wang, C. Ni, X. Shi, et al., "Cosyvoice 3: Towards in-the-wild speech generation via scaling-up and post-training," *arXiv preprint arXiv:2505.17589*, 2025.

[6] Z. Huang, C. Meng, and T. Ko, "Repcodec: A speech representation codec for speech tokenization," *arXiv preprint arXiv:2309.00169*, 2023.

[7] D. Tao, D. Tan, Y. T. Yeung, X. Chen, and T. Lee, "Toneunit: A speech discretization approach for tonal language speech synthesis," *arXiv preprint arXiv:2406.08989*, 2024.

[8] D. Xin, X. Tan, S. Takamichi, and H. Saruwatari, "Bigcodec: Pushing the limits of low-bitrate neural speech codec," *arXiv preprint arXiv:2409.05377*, 2024.

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

[10] H. Wu, N. Kanda, S. E. Eskimez, and J. Li, "Ts3-codec: Transformer-based simple streaming single codec," *arXiv preprint arXiv:2411.18803*, 2024.

[11] Y. Ren, T. Wang, J. Yi, L. Xu, J. Tao, C. Y. Zhang, and J. Zhou, "Fewer-token neural speech codec with time-invariant codes," in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2024, pp. 12 737–12 741.

[12] H. Li, L. Xue, H. Guo, X. Zhu, Y. Lv, L. Xie, Y. Chen, H. Yin, and Z. Li, "Single-codec: Single-codebook speech codec towards high-performance speech generation," *arXiv preprint arXiv:2406.07422*, 2024.

[13] Y. Zheng, W. Tu, Y. Kang, J. Chen, Y. Zhang, L. Xiao, Y. Yang, and L. Ma, "Freecodec: A disentangled neural speech codec with fewer tokens," *arXiv preprint arXiv:2412.01053*, 2024.

[14] W. Chen, X. Wang, R. Yan, Y. Chen, Z. Niu, Z. Ma, X. Li, Y. Liang, H. Wen, S. Yin, et al., "Sac: Neural speech codec with semantic-acoustic dual-stream quantization," *arXiv preprint arXiv:2510.16841*, 2025.

[15] A. Van Den Oord, O. Vinyals, et al., "Neural discrete representation learning," *Advances in neural information processing systems*, vol. 30, 2017.

[16] F. Mentzer, D. Minnen, E. Agustsson, and M. Tschannen, "Finite scalar quantization: Vq-vae made simple," *arXiv preprint arXiv:2309.15505*, 2023.

[17] C. Wang, S. Chen, Y. Wu, Z. Zhang, L. Zhou, S. Liu, Z. Chen, Y. Liu, H. Wang, J. Li, et al., "Neural codec language models are zero-shot text to speech synthesizers," *arXiv preprint arXiv:2301.02111*, 2023.

[18] A. Défossez, J. Copet, G. Synnaeve, and Y. Adi, "High fidelity neural audio compression," *Trans. Mach. Learn. Res.*, vol. 2023, 2023.

[19] J. Betker, "Better speech synthesis through scaling," *arXiv preprint arXiv:2305.07243*, 2023.

[20] E. Casanova, K. Davis, E. Gölge, G. Göknar, I. Gulea, L. Hart, A. Aljafari, J. Meyer, R. Morais, S. Olayemi, et al., "Xtts: A massively multilingual zero-shot text-to-speech model," *arXiv preprint arXiv:2406.04904*, 2024.

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

[22] Z. Du, Q. Chen, S. Zhang, K. Hu, H. Lu, Y. Yang, H. Hu, S. Zheng, Y. Gu, Z. Ma, Z. Gao, and Z. Yan, "Cosyvoice: A scalable multilingual zero-shot text-to-speech synthesizer based on supervised semantic tokens," *CoRR*, vol. abs/2407.05407, 2024.

[23] Z. Du, Y. Wang, Q. Chen, X. Shi, X. Lv, T. Zhao, Z. Gao, Y. Yang, C. Gao, H. Wang, F. Yu, H. Liu, Z. Sheng, Y. Gu, C. Deng, W. Wang, S. Zhang, Z. Yan, and J. Zhou, "Cosyvoice 2: Scalable streaming speech synthesis with large language models," *CoRR*, vol. abs/2412.10117, 2024.

[24] Y. Wang, H. Zhan, L. Liu, R. Zeng, H. Guo, J. Zheng, Q. Zhang, X. Zhang, S. Zhang, and Z. Wu, "Maskgct: Zero-shot text-to-speech with masked generative codec transformer," *arXiv preprint arXiv:2409.00750*, 2024.

[25] K. Xie, F. Shen, J. Li, F. Xie, X. Tang, and Y. Hu, "Fireredtts-2: Towards long conversational speech generation for podcast and chatbot," *arXiv preprint arXiv:2509.02020*, 2025.

[26] L. Zhang, Y. Qian, L. Zhou, S. Liu, D. Wang, X. Wang, M. Yousefi, Y. Qian, J. Li, L. He, et al., "Covomix: Advancing zero-shot speech generation for human-like multi-talker conversations," *Advances in Neural Information Processing Systems*, vol. 37, pp. 100 291–100 317, 2024.

[27] Z. Ju, D. Yang, J. Yu, K. Shen, Y. Leng, Z. Wang, X. Tan, X. Zhou, T. Qin, and X. Li, "Mooncast: High-quality zero-shot podcast generation," *arXiv preprint arXiv:2503.14345*, 2025.

[28] O. Team, "Text to spoken dialogue generation," 2025.

[29] S. Team, *Silero vad: Pre-trained enterprise-grade voice activity detector (vad), number detector and language classifier*, `https://github.com/snakers4/silero-vad`, 2024.

[30] T. Park, I. Medennikov, K. Dhawan, W. Wang, H. Huang, N. R. Koluguri, K. C. Puvvada, J. Balam, and B. Ginsburg, "Sortformer: Seamless integration of speaker diarization and asr by bridging timestamps and tokens," *arXiv preprint arXiv:2409.06656*, 2024.

[31] C. K. A. Reddy, V. Gopal, and R. Cutler, "Dnsmos P.835: A non-intrusive perceptual objective speech quality metric to evaluate noise suppressors," IEEE, 2022, pp. 886–890.

[32] Z. Chen, S. Chen, Y. Wu, Y. Qian, C. Wang, S. Liu, Y. Qian, and M. Zeng, "Large-scale self-supervised speech representation learning for automatic speaker verification," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2022, pp. 6147–6151.

[33] S. Chen, Y. Wu, C. Wang, S. Liu, D. Tompkins, Z. Chen, and F. Wei, "Beats: Audio pre-training with acoustic tokenizers," 2022. arXiv: `2212.09058`.

[34] J. Darefsky, G. Zhu, and Z. Duan, *Parakeet: A natural sounding, conversational text-to-speech model*, `https://jordandarefsky.com/blog/2024/parakeet/`, blog post, 2024.

[35] G. Comanici, E. Bieber, M. Schaekermann, I. Pasupat, N. Sachdeva, I. Dhillon, M. Blistein, O. Ram, D. Zhang, E. Rosen, et al., "Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities," *arXiv preprint arXiv:2507.06261*, 2025.

[36] Y. Chen, Z. Niu, Z. Ma, K. Deng, C. Wang, J. Zhao, K. Yu, and X. Chen, "F5-TTS: A fairytaler that fakes fluent and faithful speech with flow matching," *CoRR*, vol. abs/2410.06885, 2024.

[37] Z. Peng, J. Yu, W. Wang, Y. Chang, Y. Sun, L. Dong, Y. Zhu, W. Xu, H. Bao, Z. Wang, et al., "Vibevoice technical report," *arXiv preprint arXiv:2508.19205*, 2025.

[38] H. Zhu, W. Kang, Z. Yao, L. Guo, F. Kuang, Z. Li, W. Zhuang, L. Lin, and D. Povey, “Zipvoice: Fast and high-quality zero-shot text-to-speech with flow matching,” *arXiv preprint arXiv:2506.13053*, 2025.

[39] J. Mai, J. Ji, X. Xing, C. Yang, W. Chen, J. Xing, and X. Xu, *Mnv-17: A high-quality performative mandarin dataset for nonverbal vocalization recognition in speech*, 2025. arXiv: `2509.18196 [cs.SD]`. [Online]. Available: `https://arxiv.org/abs/2509.18196`.

[40] L. Li, Z. Guo, H. Chen, Y. Dai, Z. Zhang, H. Xue, T. Zuo, C. Wang, S. Wang, J. Li, et al., “Wenetspeech-yue: A large-scale cantonese speech corpus with multi-dimensional annotation,” *arXiv preprint arXiv:2509.03959*, 2025.

[41] Y. Dai, Z. Zhang, S. Wang, L. Li, Z. Guo, T. Zuo, S. Wang, H. Xue, C. Wang, Q. Wang, et al., “Wenetspeech-chuan: A large-scale sichuanese corpus with rich annotation for dialectal speech processing,” *arXiv preprint arXiv:2509.18004*, 2025.