STRUCTURED PROSODY MODELING FOR LONG-FORM TEXT-TO-SPEECH VIA SEMANTIC-CONDITIONED CROSS-SCALE REGULARIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Despite recent advances in neural text-to-speech synthesis, long-form generation still suffers from prosodic degradation, including inappropriate pauses, pitch drift, unstable energy patterns, and rhythmic discontinuities. We argue that these failures arise because prosody is often modeled as a flat acoustic sequence, a global style variable, or a reference-conditioned latent factor, without explicitly enforcing structural consistency across phrase-, sentence-, and discourse-level contexts. To address this limitation, we formulate long-form prosody modeling as a cross-scale structural-consistency problem and propose SPM-TTS, a structured prosody modeling framework for neural speech synthesis. SPM-TTS couples three components: Segmental Information-Guided Generation (SIG), Semantic-conditioned Prosody Consistency Regularization (S-PCR), and a neural-codec acoustic representation. Experiments on AISHELL-3 and long-form Chinese inputs show improvements in speech quality, intelligibility, speaker similarity, and prosodic stability over representative baselines. Ablation studies further indicate that segment-aware local modeling, semantic-conditioned global regularization, and codec-level acoustic representation provide complementary gains.

1 INTRODUCTION

Neural text-to-speech (TTS) systems have undergone a paradigm shift from cascaded architectures to end-to-end modeling, and their speech quality and intelligibility are approaching those of natural speech (Shen et al., 2018; Ren et al., 2019; Kim et al., 2020; Ren et al., 2021a; Kim et al., 2021; Ren et al., 2021b). Despite continued progress in acoustic fidelity, prosody control—the precise modeling of suprasegmental properties such as pitch, rhythm, and energy—remains a central bottleneck for practical TTS systems (Tan et al., 2024). This issue is especially pronounced in long-form or context-dependent generation (Jiang et al., 2024; Dang et al., 2024). In the absence of a unified mechanism for modeling across temporal scales, generated speech often exhibits unstable pitch trajectories, incoherent rhythmic patterns, and reduced overall naturalness (Chen et al., 2023). These phenomena reflect two fundamental limitations of current TTS systems: insufficient fine-grained prosody generation and inadequate long-range structural consistency (Ju et al., 2024).

From the perspective of generative modeling, prosody generation is a multi-scale structured prediction problem. At the local temporal scale, prosodic features such as fundamental frequency and energy envelopes should transition smoothly between adjacent frames. At the global temporal scale, prosodic patterns should maintain stylistic consistency and semantic coherence over an entire utterance. Existing methods, however, typically model prosody either as a flat sequence or as a global latent variable, and therefore do not explicitly characterize this hierarchical structure. Autoregressive language-model-based approaches, such as the VALL-E series (Wang et al., 2023; Chen et al., 2024a; Le et al., 2023; Wang et al., 2025a; Yang et al., 2025a; Xia et al., 2026; Cui et al., 2025; Yang et al., 2025b), acquire strong contextual modeling ability through large-scale pretraining, but their token-by-token generation mechanism is susceptible to error accumulation over long sequences and can lead to prosodic drift. Diffusion- or flow-matching-based approaches (Eskimez et al., 2024; Popov et al., 2021), such as NaturalSpeech 2 (Shen et al., 2023) and F5-TTS (Chen et al., 2024b; Mehta et al., 2024), can generate high-quality samples, but they lack explicit structural constraints and have difficulty ensuring cross-segment prosodic consistency. Style-transfer approaches, such as

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

StyleTTS 2 (Li et al., 2024b), primarily focus on global control of speaker style and offer limited modeling of local prosodic boundaries.

The choice of acoustic representation also has a substantial impact on prosody modeling. Traditional mel spectrograms or VAE-style latent representations may lose harmonic structure and temporal detail during compression (Zhang et al., 2024; Borsos et al., 2022). Neural codec tokens preserve richer acoustic information, but their discretization process can introduce quantization noise and typically lacks explicit encoding of prosodic semantics (Zeghidour et al., 2021; Baas et al., 2025; Li et al., 2024a). Designing representations that retain detailed acoustic information while supporting structured prosody control is therefore another unresolved challenge.

To address these challenges, we propose Structured Prosody Modeling for TTS (SPM-TTS). The central idea is to model prosody as a hierarchically organized signal rather than as a flat sequence. SPM-TTS integrates three mutually reinforcing components:

- **Segmental Information-Guided Generation (SIG).** By explicitly introducing linguistic boundaries, such as phrases and semantic units, SIG organizes input text into structured segments, enabling the model to learn local semantic-acoustic alignment and to distinguish smooth intra-segment transitions from appropriate inter-segment pauses (Li et al., 2025).
- **Semantic-conditioned Prosody Consistency Regularization (S-PCR).** S-PCR formulates prosody consistency as a regularization constraint in an implicit prosodic space. It suppresses long-range prosodic drift by minimizing the deviation between local prosodic features and global statistics, while adapting the constraint according to segment semantics.
- **Neural-codec acoustic representation.** A high-quality speech tokenizer encodes speech into discrete acoustic tokens. This representation preserves rich acoustic detail and supports prosody-aware generation.

The contributions of this paper are as follows. (1) We frame long-timescale prosodic drift as a cross-scale structural-consistency problem rather than merely an acoustic reconstruction problem. This formulation explicitly separates local prosodic continuity, segment-level boundary organization, and utterance-level prosodic stability. (2) We propose SPM-TTS, a structured prosody modeling framework that combines SIG with S-PCR. Unlike conventional global moment matching, S-PCR uses semantic offsets and adaptive consistency weights to preserve semantically appropriate local prosodic variation while suppressing harmful long-range prosodic drift. (3) We evaluate SPM-TTS on AISHELL-3 and long-form Chinese inputs using speech-quality, intelligibility, speaker-consistency, and pitch/energy stability metrics.

2 Related Work

2.1 Prosody Modeling

Prosody modeling is a core problem in TTS and directly affects naturalness, expressiveness, and semantic communicativeness (Xie et al., 2025). Recent studies increasingly recognize prosody generation as a one-to-many problem: the same text may admit multiple reasonable prosodic realizations under different contexts, emotions, and speaking styles (Li et al., 2024b). StyleTTS 2 models speaking style as a latent random variable and improves naturalness using diffusion and adversarial training (Li et al., 2024b). ProsodyFlow combines self-supervised speech models with conditional flow matching to model prosodic features (Huang et al., 2024). Prior work also shows that conventional mean-squared-error regression objectives can yield over-smoothed predictions for multimodal prosodic distributions (Tan et al., 2021). These methods improve prosodic diversity and naturalness, but most still rely on implicit style variables, reference speech, or global latent representations and provide limited explicit constraints on segmentation boundaries, local rhythmic structure, and cross-scale consistency.

2.2 Multi-Speaker Synthesis and Prosody Transfer

Multi-speaker speech synthesis and prosody transfer are important directions for improving TTS controllability. CopyCat2 supports multi-speaker synthesis, context-dependent prosody generation, and fine-grained prosody transfer (Zhang et al., 2023). eCat further targets multi-speaker TTS and

many-to-many fine-grained prosody transfer (Kim et al., 2023). These approaches show that speaker identity, timbre, and prosodic style are intricately coupled. However, existing prosody-transfer methods usually focus on reference-driven style reproduction and pay less attention to long-range prosodic consistency under no-reference or weak-reference conditions. This paper starts from the structure of the text itself: explicit segmentation information constrains the local generation process, while S-PCR constrains prosodic coherence across segments.

2.3 Speech Tokenizers and Acoustic Representations

Discrete speech tokens and codec representations have become important foundations of modern TTS systems. VALL-E formulates TTS as conditional language modeling over neural audio codec tokens (Wang et al., 2023), while VALL-E 2 introduces repetition-aware sampling and grouped code modeling (Chen et al., 2024a). SpeechTokenizer (Zhang et al., 2024) adopts an encoder-decoder architecture with residual vector quantization, hierarchically disentangling semantic and acoustic information across quantization layers. Neural audio codecs such as SoundStream (Zeghidour et al., 2021) preserve compact acoustic information for waveform reconstruction. Recent speech-generation systems further explore codec and tokenized representations, including MaskGCT (Wang et al., 2025b), RALL-E (Xin et al., 2024), token-transducer-based synthesis (Lee et al., 2024), Seed-TTS (Anastassiou et al., 2024), HierSpeech++ (Lee et al., 2023), and continuous-tokenizer approaches (Huang et al., 2025; Meng et al., 2025; Liu et al., 2026). Powerful representations, however, do not by themselves guarantee structured prosody control. Our method therefore adds segmentation structure and prosody-consistency constraints on top of codec-based acoustic modeling.

3 Method

3.1 Problem Formulation

We formalize neural speech generation as a conditional generation task that maps a text sequence $X = (x_1, x_2, \ldots, x_T)$ to a speech signal $Y = (y_1, y_2, \ldots, y_S)$. Prosody is represented as a structured signal $P = (p_1, p_2, \ldots, p_S)$ containing fundamental frequency (F_0), short-time energy, and duration patterns. These attributes have two basic properties. First, within short temporal windows, adjacent-frame prosodic features should satisfy a smooth-transition constraint, $P_t \approx P_{t+1}$. Second, at the utterance level, local prosodic patterns should remain consistent with global statistics, $P_i \sim P_{\text{global}}$. Existing methods usually model these aspects separately, leading to a mismatch between local generation and global structure.

3.2 Structured Prosody Modeling Framework

To address the above limitations, we propose Structured Prosody Modeling (SPM), a unified framework that explicitly enforces consistency across temporal scales. The framework decomposes prosody modeling into local structure modeling through SIG, global consistency modeling through S-PCR, and representation learning through a neural codec. These components are integrated into

$$\mathcal{L} = \mathcal{L}_{\text{rec}} + \lambda \mathcal{L}_{\text{S-PCR}}, \tag{1}$$

where $\mathcal{L}_{\text{rec}}$ is the reconstruction loss, $\mathcal{L}_{\text{S-PCR}}$ is the prosody-consistency constraint, and λ controls the regularization strength.

3.3 Training and Inference Overview

During training, SPM-TTS receives an input text sequence and its corresponding speech target. SIG first converts the flat text sequence into segment-aware textual units. The acoustic generator then predicts codec-level acoustic representations conditioned on the segment-aware text representation. In parallel, S-PCR is applied as a training-time regularizer: it extracts segment-level prosodic statistics from the predicted acoustic representation and aligns them with a semantic-conditioned global prosodic target. This encourages the model to learn both local boundary-aware generation and utterance-level prosodic stability. During inference, only the text segmentation and acoustic generation pathways are used. S-PCR is removed because it serves as a training-time structural constraint, so it does not introduce additional iterative decoding or external prosody prediction at inference time.

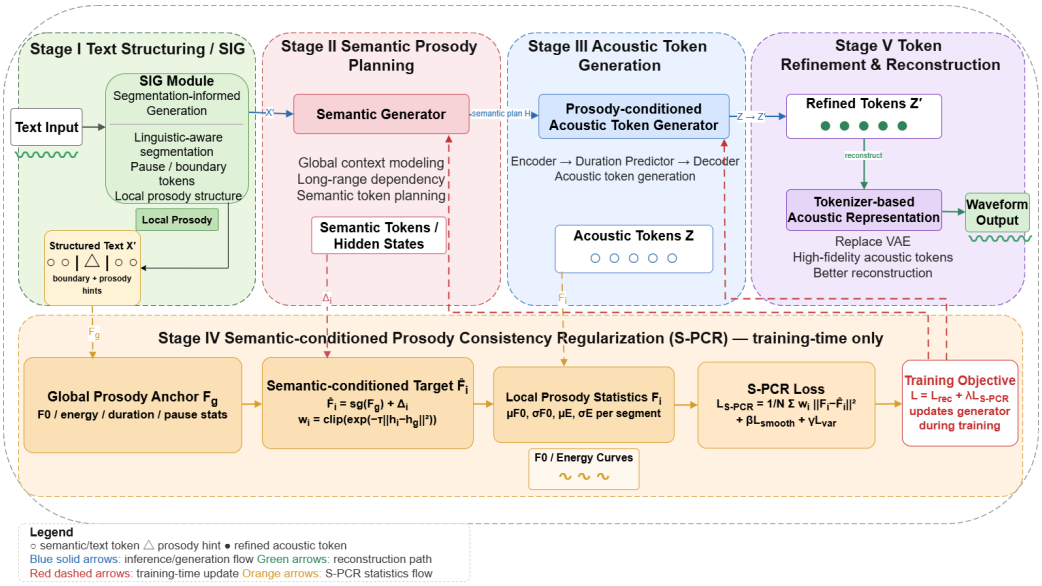

Figure 1: Overview of the proposed Structured Prosody Modeling framework in SPM-TTS. SIG enhances local temporal consistency through linguistically motivated segmentation, while S-PCR enforces global alignment between local prosodic patterns and sentence-level structure. SpeechTokenizer provides the acoustic representation used for reconstruction.

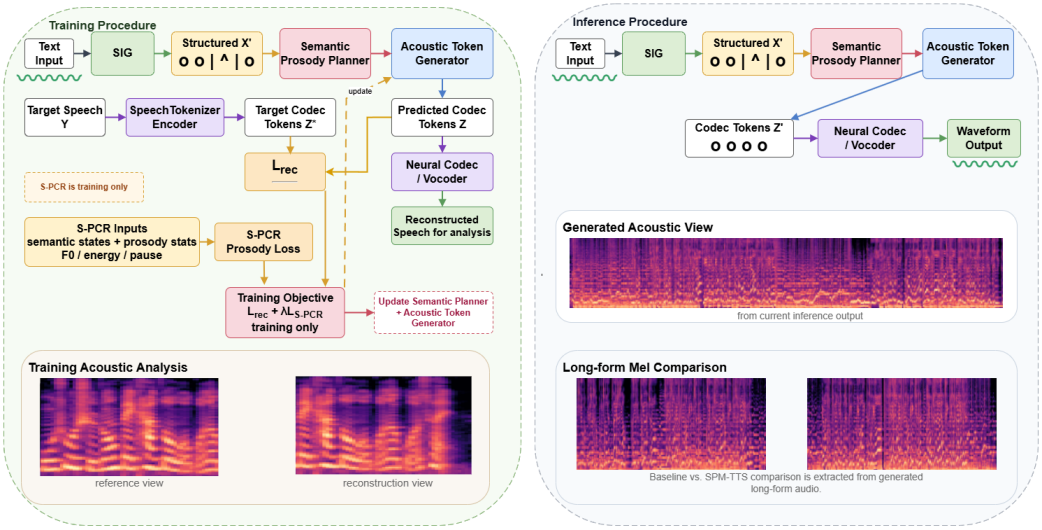

Figure 2: Training and inference workflow of SPM-TTS. S-PCR is applied only during training, while inference uses text segmentation, acoustic token generation, and neural codec reconstruction.

3.4 Segmental Information-Guided Generation

Standard TTS models treat the input text as a flat sequence and ignore natural linguistic boundaries such as phrases or semantic units. SIG addresses this problem by explicitly introducing segmental information. Given an input text sequence X, we organize it into N semantically meaningful segments,

$$X = \{X_1, X_2, \ldots, X_N\}, \tag{2}$$

where each X_i corresponds to a semantic unit such as a phrase, clause, or intonational phrase. In practice, segmentation is implemented by (1) inserting special boundary tokens such as $\langle \mathrm{SEG} \rangle$; (2) introducing segmental position encodings into the attention mechanism to distinguish intra-segment and inter-segment relations (Lu et al., 2025); and (3) conditioning acoustic prediction on segment information. Unlike conventional fixed-patch interleaving, SIG introduces dynamically determined linguistic boundaries (Zhang et al., 2025) and can adaptively group semantic and acoustic tokens.

3.5 Semantic-Conditioned Prosody Consistency Regularization

Although SIG improves local prosodic structure through linguistic boundaries, long-form generation may still suffer from cross-segment prosodic drift. Simple global mean matching ignores local semantic differences: questions, contrastive phrases, and emphasized segments should naturally have distinct pitch, energy, or pause patterns. We therefore propose S-PCR, which aims to preserve global prosodic stability while allowing reasonable local offsets caused by semantic variation.

Given text segments $\{X_i\}_{i=1}^{N}$, we extract prosodic statistics for the i-th segment:

$$F_i = [\mu_i^{F0}, \sigma_i^{F0}, \mu_i^{E}, \sigma_i^{E}], \tag{3}$$

where $F0$ and E denote fundamental frequency and energy, and μ and σ denote within-segment mean and standard deviation. The corresponding global prosodic statistics are denoted by F_g. Segment-level and global semantic representations are obtained using a semantic encoder:

$$h_i = \mathrm{Pool}(\mathrm{Enc}(X_i)), \qquad h_g = \mathrm{Pool}(\mathrm{Enc}(X)). \tag{4}$$

To avoid forcing all segments toward the same global mean, segment semantics predict an allowable offset:

$$\delta_i = \mathrm{MLP}([h_i; h_g; h_i - h_g]). \tag{5}$$

The semantic-conditioned target is

$$\hat{F}_i = \mathrm{sg}(F_g) + \delta_i, \tag{6}$$

and the adaptive consistency weight is

$$w_i = \mathrm{clip}(\exp(-\tau \|h_i - h_g\|_2), w_{\min}, w_{\max}). \tag{7}$$

Segments semantically closer to the global context receive stronger consistency constraints, whereas semantic turns or emphasized segments retain greater local freedom. The final regularization loss is

$$\mathcal{L}_{\text{S-PCR}} = \frac{1}{N} \sum_{i=1}^{N} w_i \|F_i - \hat{F}_i\|_2^2 + \beta \mathcal{L}_{\mathrm{smooth}} + \gamma \mathcal{L}_{\mathrm{var}}. \tag{8}$$

Here, $\mathcal{L}_{\mathrm{smooth}}$ constrains second-order differences of frame-level F_0 and energy trajectories, while $\mathcal{L}_{\mathrm{var}}$ preserves necessary inter-segment prosodic variance. The overall training objective is Eq. 1. Unlike ordinary prosody-consistency regularization that directly minimizes $\|F_i - F_g\|_2^2$, S-PCR extends the global anchor into the semantic-conditioned target $\hat{F}_i$.

3.6 Neural-Codec Acoustic Representation

We use SpeechTokenizer to encode speech into discrete acoustic tokens with residual vector quantization (Zhang et al., 2024). Compared with conventional VAE-style latent representations, tokenizer-based acoustic representations preserve fine-grained acoustic detail and provide a compact space for prosody-aware generation. This representation is compatible with segment-conditioned acoustic prediction, allowing local boundary information introduced by SIG to be reflected in codec-level generation.

4 Experiments

4.1 Experimental Setup

We evaluate SPM-TTS on AISHELL-3, a Mandarin multi-speaker TTS corpus containing approximately 85 hours of speech from 218 speakers (Shi et al., 2021). Mandarin long-form synthesis requires

Table 1: Objective evaluation results, part I. Higher is better.

Model	PESQ ↑	STOI ↑	F_0 corr. ↑
Baseline	3.85	0.96	0.75
F5-TTS	3.92	0.97	0.82
CosyVoice3	3.95	0.97	0.80
Qwen3-TTS	3.88	0.97	0.78
SPM-TTS	**4.21**	**0.99**	**0.86**

Table 2: Objective evaluation results, part II. Lower WER is better; higher T-SIM, S-SIM, and DNSMOS are better.

Model	WER ↓	T-SIM ↑	S-SIM ↑	DNSMOS ↑
F5-TTS	2.05	0.74	–	3.91
CosyVoice3	1.22	0.74	0.44	3.92
Qwen3-TTS	1.03	0.65	0.33	3.90
SPM-TTS	**0.94**	**0.75**	**0.62**	**3.95**

stable modeling of tone, rhythm, pauses, and syntactic-semantic boundaries, making AISHELL-3 and held-out Chinese long-form inputs suitable for evaluating cross-scale prosody consistency. All objective results are computed on the official AISHELL-3 test split. The same text inputs, reference prompts, and evaluation scripts are used for all compared systems.

We compare SPM-TTS with a fixed-block AR-Diffusion baseline and representative recent TTS systems, including F5-TTS (Chen et al., 2024b), CosyVoice3 (Du et al., 2025), and Qwen3-TTS (Hu et al., 2026). Since large external systems may differ in training data, model scale, and pretraining strategy, they are treated as representative reference baselines rather than strictly training-budget-matched systems. The reproduced baseline takes structured text and a reference audio prompt as input. The generated waveform is saved in FLAC format at 48 kHz. On a single NVIDIA Tesla V100S-PCIE-32GB GPU, the baseline requires 126.5 seconds per sample and approximately 27.5 GB peak GPU memory. The main sampling parameters are $\mathrm{cfg_coef} = 1.5$, steps$= 36$, top-k$= 100$, temperature$= 0.9$, and diffusion temperature$= 0.95$.

4.2 Evaluation Metrics

We use speech-quality, intelligibility, speaker-consistency, and prosody-related metrics. PESQ (Rix et al., 2001) and DNSMOS (Reddy et al., 2021) are used as speech-quality indicators, while STOI (Taal et al., 2011) and WER evaluate intelligibility and content preservation. T-SIM and S-SIM evaluate speaker consistency using the same speaker-embedding model for all systems. F_0 correlation, F_0 RMSE, and Energy RMSE measure prosodic stability. Since objective metrics cannot fully capture perceived pause naturalness, rhythm continuity, and discourse-level expressiveness, we further conduct a mean opinion score (MOS) listening test for perceived naturalness. Full implementation details are provided in Appendix B.

4.3 Main Results

As shown in Tables 1 and 2, SPM-TTS achieves favorable objective results across speech quality, intelligibility, speaker consistency, and prosody-related metrics under a unified evaluation protocol. PESQ improves from 3.85 for the reproduced baseline to 4.21, STOI from 0.96 to 0.99, and F_0 correlation from 0.75 to 0.86. The S-SIM result also indicates reduced speaker drift under the adopted speaker-embedding metric.

4.4 Ablation Study

Table 3 reports the ablation results on AISHELL-3. The results show a consistent upward trend across objective quality, intelligibility, prosody-stability, and human-rated naturalness metrics. SIG

Table 3: Ablation study on AISHELL-3. Higher is better.

Variant	PESQ ↑	STOI ↑	F_0 corr. ↑	MOS ↑
Baseline	3.85	0.96	0.75	4.00
+ SIG	3.92	0.97	0.80	4.20
+ S-PCR	4.05	0.98	0.83	4.30
+ Tokenizer	4.15	0.98	0.85	4.35
Full model	**4.21**	**0.99**	**0.86**	**4.40**

Table 4: Long-form generation results. Lower is better.

Model	WER ↓	F_0 RMSE ↓	Energy RMSE ↓
Baseline	3.42	18.5	0.28
F5-TTS	2.87	15.2	0.24
CosyVoice3	2.15	12.8	0.21
Qwen3-TTS	1.89	11.5	0.19
SPM-TTS	**1.23**	**8.7**	**0.15**

improves F_0 correlation from 0.75 to 0.80. Adding S-PCR further improves PESQ, F_0 correlation, and MOS. The tokenizer-based acoustic representation provides additional gains in the full system.

4.5 Long-Form Evaluation

Since the primary goal of this study is long-form prosody consistency, we further evaluate the models on Chinese inputs of 200–2000 words. Table 4 shows the results. SPM-TTS reduces WER, F_0 RMSE, and Energy RMSE compared with the evaluated systems under the reported protocol.

5 Conclusion and Limitations

We presented SPM-TTS, a structured prosody modeling framework that treats long-form TTS as a cross-scale consistency problem. By combining SIG, S-PCR, and neural-codec acoustic representation, SPM-TTS improves local boundary modeling, semantic-conditioned global prosody stability, and codec-level acoustic generation. Experiments on AISHELL-3 and long-form Chinese inputs show improvements in speech quality, intelligibility, speaker consistency, and pitch/energy stability under the reported evaluation protocol. The method still depends on segmentation quality and is currently evaluated mainly on Mandarin speech. Automatic metrics also cannot fully capture discourse-level expressiveness. Future work will extend SPM-TTS to multilingual, emotional, and more diverse long-form synthesis settings.

AI use statement

Generative AI tools were used for language editing, LaTeX formatting, and manuscript organization assistance. They were not used to generate experimental data, train the models, or create the reported experimental results. All AI-assisted edits were reviewed by the authors, who take responsibility for the final text, claims, and artifacts in this submission.

Ethics statement

This work studies neural speech synthesis for academic research. TTS systems may be misused for unauthorized voice replication, impersonation, deceptive audio generation, misinformation, harassment, or privacy violations. Speaker references should be used with appropriate consent or licensing, and generated speech should be disclosed as synthetic outside controlled research settings. The MOS study used anonymized synthetic speech samples and collected no personally identifiable information; institutional requirements were handled according to the authors' local policy.

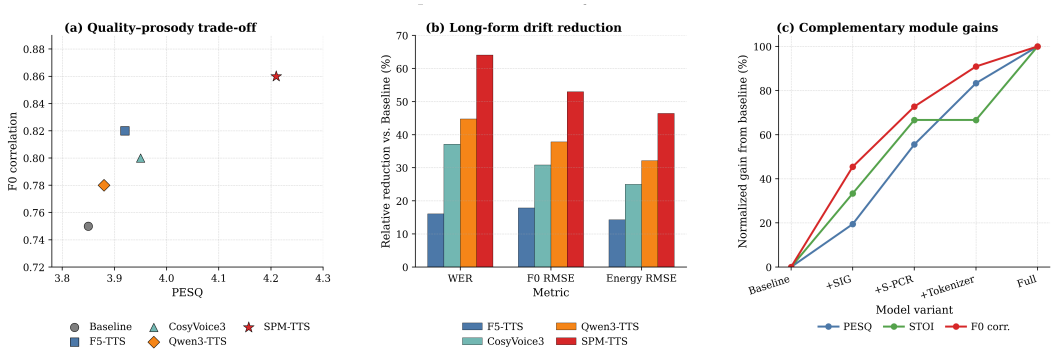

Figure 3: Core experimental summary of SPM-TTS. Left: quality–prosody trade-off using PESQ and F_0 correlation. Middle: relative reductions in WER, F_0 RMSE, and Energy RMSE with respect to the fixed-block baseline. Right: normalized ablation gains from Baseline to Full model.

Reproducibility Statement

Implementation and evaluation details are provided in Appendices A and B, including baseline inference settings, the SIG static validation, metric definitions, long-form evaluation settings, and subjective-evaluation procedures. The paper cites the public datasets and external systems used in the study. Full training code, model checkpoints, and processed data are not released at submission time.

References

Philip Anastassiou et al. Seed-tts: A family of high-quality versatile speech generation models. *arXiv preprint arXiv:2406.02430*, 2024.

Matthew Baas et al. Mars6: A small and robust hierarchical-codec text-to-speech model. In *ICASSP*, 2025.

Zalán Borsos et al. Audiolm: A language modeling approach to audio generation. *arXiv preprint arXiv:2209.03143*, 2022.

M. Chen et al. Long-form text-to-speech: A survey. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2023.

Sanyuan Chen et al. Vall-e 2: Neural codec language models are human parity zero-shot text to speech synthesizers. *arXiv preprint arXiv:2406.05370*, 2024a.

Yushen Chen et al. F5-tts: A fairytaler that fakes fluent and faithful speech with flow matching. *arXiv preprint arXiv:2410.06885*, 2024b.

Wenqian Cui et al. Recent advances in speech language models: A survey. In *ACL*, 2025.

Tu Anh Dang et al. Zero-shot text-to-speech from continuous text streams. *arXiv preprint arXiv:2410.00767*, 2024.

Zhihao Du et al. Cosyvoice 3: Towards in-the-wild speech generation via scaling-up and post-training. *arXiv preprint arXiv:2505.17589*, 2025.

Sefik Emre Eskimez et al. E2 tts: Embarrassingly easy fully non-autoregressive zero-shot tts. *arXiv preprint arXiv:2406.18009*, 2024.

Hangrui Hu et al. Qwen3-tts technical report. *arXiv preprint arXiv:2601.15621*, 2026.

H. Huang et al. Continuous speech tokenizers for tts and audio generation. In *ICASSP*, 2025.

R. Huang et al. Prosodyflow: Learning prosody from speech via flow matching. In *ICASSP*, 2024.

Ziyue Jiang et al. Mega-tts 2: Zero-shot text-to-speech with arbitrary length speech prompts. In *ICLR*, 2024.

Zeqian Ju et al. Naturalspeech 3: Zero-shot speech synthesis with factorized codec and diffusion models. *arXiv preprint arXiv:2403.03100*, 2024.

J. Kim et al. ecat: End-to-end fine-grained prosody transfer for tts. In *Interspeech*, 2023.

Jaehyeon Kim, Sungwon Kim, Jungil Kong, and Sungroh Yoon. Glow-tts: A generative flow for text-to-speech via monotonic alignment search. In *NeurIPS*, 2020.

Jaehyeon Kim, Jungil Kong, and Juhee Son. Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech. In *ICML*, 2021.

Matthew Le et al. Voicebox: Text-guided multilingual universal speech generation at scale. In *NeurIPS*, 2023.

Joun Yeop Lee et al. High fidelity text-to-speech via discrete tokens using token transducer and group masked language model. In *Interspeech*, 2024.

Sang-Hoon Lee, Ha-Yeong Choi, Seung-Bin Kim, and Seong-Whan Lee. Hierspeech++: Bridging the gap between semantic and acoustic representation of speech by hierarchical variational inference for zero-shot speech synthesis. *arXiv preprint arXiv:2311.12454*, 2023.

Jiaqi Li et al. Investigating neural audio codecs for speech language model-based speech generation. *arXiv preprint arXiv:2409.04016*, 2024a.

Xiaofeng Li et al. Muyan-tts: A trainable text-to-speech model optimized for podcast generation. *arXiv preprint arXiv:2504.19146*, 2025.

Yinghao Li et al. Styletts 2: Towards human-level text-to-speech through style diffusion and adversarial training. In *ICLR*, 2024b.

L. Liu et al. Dashengtokenizer: A unified continuous audio tokenizer for speech, music, and audio. *arXiv preprint arXiv:2602.23765*, 2026.

Chunhui Lu, Xue Wen, Liming Song, and Junkwang Oh. Robust neural codec language modeling with phoneme position prediction for zero-shot tts. In *Interspeech*, 2025.

Shivam Mehta et al. Matcha-tts: A fast tts architecture with conditional flow matching. In *ICASSP*, 2024.

Lingwei Meng et al. Melle: Autoregressive speech synthesis without vector quantization. In *ACL*, 2025.

Vadim Popov et al. Grad-tts: A diffusion probabilistic model for text-to-speech. In *ICML*, 2021.

Chandan K. A. Reddy, Vishak Gopal, and Ross Cutler. Dnsmos p.835: A non-intrusive perceptual objective speech quality metric to evaluate noise suppressors. *arXiv preprint arXiv:2110.01763*, 2021.

Yi Ren, Yangjun Ruan, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, and Tie-Yan Liu. Fastspeech: Fast, robust and controllable text to speech. In *NeurIPS*, 2019.

Yi Ren et al. Fastspeech 2: Fast and high-quality end-to-end text to speech. In *ICLR*, 2021a.

Yi Ren et al. Portaspeech: Portable and high-quality generative text-to-speech. In *NeurIPS*, 2021b.

Antony W. Rix, John G. Beerends, Michael P. Hollier, and Andries P. Hekstra. Perceptual evaluation of speech quality (pesq): A new method for speech quality assessment of telephone networks and codecs. In *ICASSP*, 2001.

Jonathan Shen, Ruoming Pang, Ron J. Weiss, Mike Schuster, Navdeep Jaitly, Zongheng Yang, Zhifeng Chen, Yu Zhang, Yuxuan Wang, R. J. Skerry-Ryan, R. A. Saurous, Yannis Agiomyrgiannakis, and Yonghui Wu. Natural tts synthesis by conditioning wavenet on mel spectrogram predictions. In *ICASSP*, 2018.

Kai Shen et al. Naturalspeech 2: Latent diffusion models are natural and zero-shot speech and singing synthesizers. In *ICLR*, 2023.

Yao Shi, Hui Bu, Xin Xu, Shaoji Zhang, and Ming Li. Aishell-3: A multi-speaker mandarin tts corpus. In *Interspeech*, pp. 2756–2760, 2021.

Cees H. Taal, Richard C. Hendriks, Richard Heusdens, and Jesper Jensen. An algorithm for intelligibility prediction of time-frequency weighted noisy speech. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(7):2125–2136, 2011.

Xu Tan et al. Prosody-tts: An end-to-end prosody-enhanced tts. In *NeurIPS*, 2021.

Xu Tan et al. Naturalspeech: End-to-end text-to-speech synthesis with human-level quality. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

Chengyi Wang et al. Vall-e: Neural codec language models are zero-shot text to speech synthesizers. *arXiv preprint arXiv:2301.02111*, 2023.

Xin Wang et al. Spark-tts: An efficient llm-based text-to-speech model with single-stream decoupled speech tokens. *arXiv preprint arXiv:2503.01710*, 2025a.

Yuancheng Wang et al. Maskgct: Zero-shot text-to-speech with masked generative codec transformer. In *ICLR*, 2025b.

Kangxiang Xia et al. Kall-e: Autoregressive speech synthesis with next-distribution prediction. In *AAAI*, 2026.

Tianxin Xie, Yan Rong, Pengfei Zhang, and Li Liu. Towards controllable speech synthesis in the era of large language models: A survey. In *EMNLP*, 2025.

Detai Xin et al. Rall-e: Robust codec language modeling with chain-of-thought prompting for text-to-speech synthesis. *arXiv preprint arXiv:2404.03204*, 2024.

Yifan Yang et al. Pseudo-autoregressive neural codec language models for efficient zero-shot text-to-speech synthesis. *arXiv preprint arXiv:2504.10352*, 2025a.

Zhongyuan Yang et al. When large language models meet speech: A survey on integration approaches. In *Findings of ACL*, 2025b.

Neil Zeghidour et al. Soundstream: An end-to-end neural audio codec. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30:495–507, 2021.

Xin Zhang et al. Speechtokenizer: Unified speech tokenizer for speech language models. In *ICLR*, 2024.

Y. Zhang et al. Copycat2: A unified model for multi-speaker tts and prosody transfer. In *ICASSP*, 2023.

Yike Zhang et al. Monotonic attention for robust text-to-speech synthesis in large language model frameworks. In *Interspeech*, 2025.

A Implementation Details

A.1 Baseline Reproduction

We reproduce the fixed-block AR-Diffusion baseline under the same inference and evaluation protocol used for SPM-TTS. The reproduced baseline takes structured text and a reference audio prompt as input. The text is converted into phoneme-level representations, while the reference audio is used as acoustic conditioning. The generated waveform is saved in FLAC format at 48 kHz. Inference is performed on a single NVIDIA Tesla V100S-PCIE-32GB GPU. Table 5 reports the reproduction configuration.

Table 5: Reproduction configuration of the fixed-block AR-Diffusion baseline.

Item	Configuration
Input condition	Structured text + reference audio prompt
Text representation	Phoneme-level representation
Output format	FLAC
Sampling rate	48 kHz
GPU	NVIDIA Tesla V100S-PCIE-32GB
Inference latency	126.5 s / sample
Peak GPU memory	27.5 GB
GPU utilization	97%
CFG coefficient	1.5
Sampling steps	36
Top-k	100
Temperature	0.9
Diffusion temperature	0.95

A.2 SIG Static Validation

We conduct a static validation of the SIG preprocessing module to verify that dynamic segmentation preserves structural markers in the input text. This validation is used only as an implementation sanity check and not as primary evidence for the main performance claims. The broken-tag rate is 0.0 over the two checked long-form samples; the average, minimum, and maximum segment counts are 52.5, 40, and 65, respectively.

A.3 Training-time Regularization and Inference-time Computation

S-PCR is used only as a training-time regularizer. During inference, the model uses the text-segmentation path and the acoustic-token generation path. The S-PCR loss is removed after training and does not introduce additional iterative decoding, auxiliary prosody prediction, external optimization, or reference-time prosody correction.

B Evaluation Protocol

B.1 Evaluation Scope

All compared systems are evaluated using the same text inputs and the same metric-computation pipeline. External TTS systems are treated as representative reference baselines rather than strictly training-budget-matched baselines because their training data, model scale, pretraining strategy, and post-training procedures may differ from those of the proposed system.

B.2 Objective Metrics

WER is computed by transcribing generated speech with the same Mandarin ASR pipeline and comparing the transcription with the target text under the same text-normalization rule. F_0 and energy features are extracted using the same frame-level analysis procedure for all evaluated systems. F_0-based metrics are computed on valid voiced regions, and energy-based metrics are computed consistently across generated samples.

B.3 Subjective Listening Evaluation

A total of 24 native Mandarin listeners participate in the MOS test. Each listener rates 25 randomly selected samples covering five model variants with the same text inputs, for 600 ratings in total. Generated samples are anonymized and randomly shuffled, and listeners are not informed of the model identity. Ratings use a five-point naturalness scale. Confidence intervals and formal statistical significance tests are not reported in the current version.

Table 6: Metric implementation details.

Metric	Implementation detail
WER	FunASR Paraformer-large Mandarin ASR; generated speech is resampled to 16 kHz and punctuation is removed before WER computation.
F_0	WORLD/pyworld DIO + StoneMask, 5 ms frame period, voiced-frame masking.
Energy	RMS energy on 48 kHz waveforms, frame length 2048, hop length 512.
PESQ	Standard PESQ implementation under matched sampling rate.
STOI	Standard STOI implementation under matched sampling rate.
DNSMOS	DNSMOS P.835 implementation.
T-SIM/S-SIM	Same ECAPA-TDNN speaker-embedding model for all systems.

B.4 Metric Definitions

Let $\hat{f}_t$ and f_t denote generated and reference F_0 values at frame t, and let V denote the set of valid voiced frames. We compute

$$\mathrm{RMSE}_{F0} = \sqrt{\frac{1}{|V|}\sum_{t\in V}(\hat{f}_t - f_t)^2}, \tag{9}$$

$$\rho_{F0} = \mathrm{corr}(\{\hat{f}_t\}_{t\in V}, \{f_t\}_{t\in V}), \tag{10}$$

$$\mathrm{RMSE}_{E} = \sqrt{\frac{1}{T}\sum_{t=1}^{T}(\hat{e}_t - e_t)^2}. \tag{11}$$

These definitions are applied consistently across evaluated systems.

B.5 Long-form Evaluation Setting

The long-form evaluation set contains 40 Chinese passages grouped into four word-count ranges: 200–500, 500–1000, 1000–1500, and 1500–2000 words, with 10 passages in each range. The passages are held out from training and manually checked to preserve coherent discourse structure rather than isolated sentence concatenation. All systems synthesize the same long-form inputs and are evaluated under the same objective metric pipeline.

C Claim-to-Evidence Map and Interpretation Boundaries

The core claims of the paper are supported by the main objective results, the ablation study, and the long-form evaluation. SPM-TTS is not presented as a new speech tokenizer; the tokenizer is used as an acoustic representation. Automatic F_0 or energy metrics do not fully capture human perception of naturalness, pause placement, or discourse-level expressiveness. External reference systems are not treated as strictly training-budget-matched baselines because their training resources and post-training procedures may differ. The paper also does not claim multilingual or emotional generalization.

D Additional Limitations and Ethical Considerations

SPM-TTS depends on the quality of text segmentation. Ambiguous discourse structures, irregular punctuation, highly conversational expressions, or expressive styles underrepresented in training data may yield suboptimal boundaries. The current evaluation is mainly conducted on Mandarin, and automatic metrics cannot fully replace human perception of pause naturalness, rhythm continuity, discourse-level expressiveness, and speaking-style appropriateness. Future work will include broader languages and styles, larger-scale human preference tests, and finer-grained boundary-level prosody metrics.