AUDIOBOOK-CC: CONTROLLABLE LONG-CONTEXT SPEECH GENERATION FOR MULTICAST AUDIOBOOK

Min Liu, JingJing Yin, Xiang Zhang, Siyu Hao, Yanni Hu, Bin Lin, Yuan Feng, Hongbin Zhou, Jianhao Ye⋆

Ximalaya Inc., China
{min1.liu, jingjing.yin, xiang2.zhang, siyu.hao, yanni.hu,bin.lin,yuan.feng,
hongbin.zhou,jianhao.ye}@ximalaya.com

ABSTRACT

Existing text-to-speech systems predominantly focus on single-sentence synthesis and lack adequate contextual modeling as well as fine-grained performance control capabilities for generating coherent multicast audiobooks. To address these limitations, we propose a context-aware and emotion controllable speech synthesis framework specifically engineered for multicast audiobooks with three key innovations: a context mechanism for contextual consistency, a disentanglement paradigm to decouple style control from speech prompts for semantic consistency, and self-distillation to boost emotional expressiveness and instruction controllability. Experimental results show superior performance across the generation of narration, dialogue, and the whole chapter, significantly outperforming existing baselines. Ablation studies are conducted to validate the effectiveness of our proposed methods. Demo samples can be found in https://everest-ai.github.io/.

Index Terms— Audiobook Generation, Context-aware, Style Control, Emotional TTS

1. INTRODUCTION

Long-form audiobooks, a widely adopted content format, integrate information delivery and auditory engagement. However, traditional production methods, whether fully manual or human-involved, face high costs and long production cycles, especially for multicast albums with multiple characters.

In response, researchers have recently developed automated solutions for high-quality audiobook generation [1–5]. AudioStory [2] employs an LLM to process instruction inputs, decomposes long audio into structured subtasks, and generates short clips sequentially. MultiActor-Audiobook [1] utilizes a Transformer-based multimodal model to capture character traits, uses LLMs for emotional guidance, and synthesizes speech at the sentence level. Dopamine Audiobook [3] and MM-StoryAgent [4] adopt multi-agent pipelines for story-based generation, but integrate existing TTS systems such as CosyVoice [6] rather than proposing new synthesis methods. Similarly, Shaja et al. [5] proposed a system enhancing immersion via spatial audio, which also relies on established TTS backbones. A key limitation across these methods is the lack of explicit inter-sentence modeling, resulting in inadequate contextual consistency.

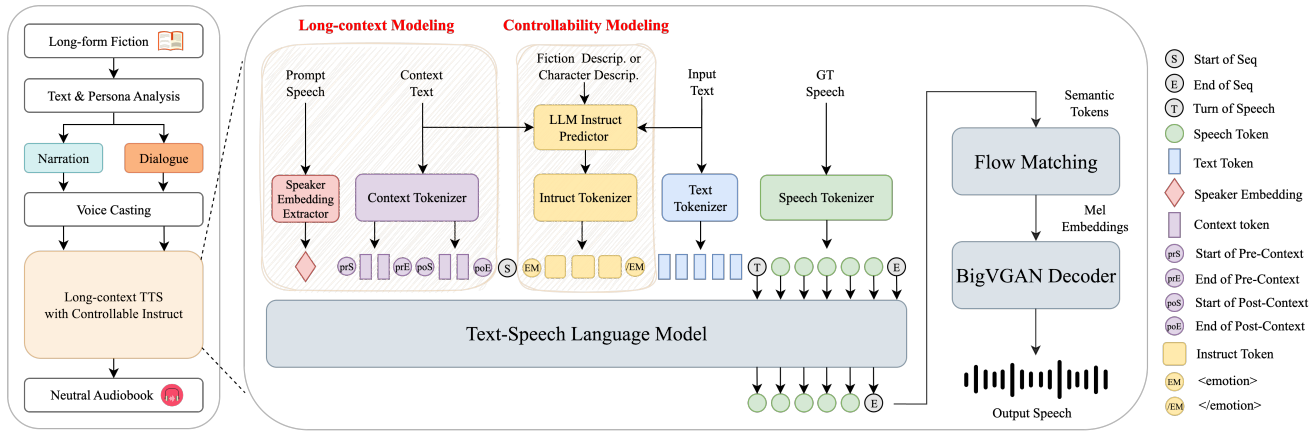
Several industry approaches have incorporated context modeling for long-form speech. MoonCast [7] targets podcast generation with long-context modeling and colloquial scripts; MOSS-TTSD [8] achieves state-of-the-art long-segment quality via efficient codecs and data pipelines; CoVoMix [9] and koel-TTS [10] focus on conversational expressiveness. Despite these advances, such systems remain largely tailored to podcasts, modeling long context in an simplistic manner that lacks fine-grained controllability—a critical requirement for audiobooks, which demand precise narrative flow and expressive multi-character portrayal.

Some prior efforts partially address context or controllability but exhibit inherent shortcomings. The prosody analysis in [11] offers data without synthesis capabilities; TACA-TTS [12] improves long-sequence continuity but lacks emotional control; the memory module in [13] over-relies on extended context, which we find can introduce redundancy and persona inconsistency. CosyVoice 2 [6] uses instructed data to improve controllability, yet still falls short in scene adaptation for audiobooks.

To overcome these limitations, we propose a novel speech synthesis framework tailored for long-form multicast audiobooks named Audiobook-CC. Our main contributions include:

- We introduce a context modeling mechanism for long-form audiobook generation, which dramatically improves semantic consistency across context.
- A disentanglement training paradigm is designed to decouple the style of the generated speech from speech prompt, which facilitates the tone of generated speech following more the semantic information of given text.

⋆Corresponding author.



(a) Audiobook Generation Pipeline (b) System Framework for Controllable Long-context Speech Generation

Fig. 1. An overview of Audiobook-CC. (a) The audiobook generation pipeline: long-form fiction is segmented into chapters, followed by textual and character persona analysis. Narrations and dialogues are extracted, assigned to voices through casting, and synthesized using the proposed architecture, resulting in a neutral-style multicast audiobook. (b) Detailed architecture of the proposed controllable long-context speech synthesis model.

- A self-distillation method is proposed to highly enhance the emotional expressiveness and instruction controllability of generated speech.

2. METHODOLOGY

We address four key challenges in long-form multicast audiobook synthesis: contextual scenario adaptability, semantic-prosodic consistency, character persona stability and controllability.To tackle these challenges, we propose the Audiobook-CC system, an integrated audiobook generation pipeline illustrated in Fig. 1(a).

The main architecture of our proposed speech generation model shown in Fig.1(b) leverages CosyVoice2 [6]. We preserve its text&speech tokenizer and flow-matching modules, but notably substitute the HiFi-GAN vocoder with BigVGAN to enhance audio fidelity and robustness. Our work mainly focuses on exploring how to better utilize contextual information and enhance controllability in the training process of large language models. For auto-regressive LLMs, the organization of sequential structures is of critical importance. In our system, it is constructed as follows:

$$[V, seqC, \textcircled{S}, seqE, \{W^i\}, \textcircled{T}, \{U^j\}, \textcircled{E}] \tag{1}$$

The symbols $\textcircled{S}$ and $\textcircled{E}$ denote the start and end of a sequence, while $\textcircled{T}$ marks the switch between text tokens and speech tokens. The text tokens W^i are obtained by processing the input text with a text tokenizer. Similarly, the speech tokens U^j are derived using the same speech tokenizer in cosyvoice2.

A speaker embedding V, derived via Cam++ [14], is prepended to the input. Contextual information is structured into pre-context C_{pre} and post-context C_{post}, demarcated by $\textcircled{preS}$/$\textcircled{preE}$ and $\textcircled{poS}$/$\textcircled{poE}$, respectively. The full contextual sequence is formed as:

$$seqC = \textcircled{preS}, \{C^m_{pre}\}, \textcircled{preE}, \textcircled{poS}, \{C^n_{post}\}, \textcircled{poE} \tag{2}$$

Control inputs (e.g., emotion or volume) are denoted E^l, with $\textcircled{EM}$ and $\textcircled{/EM}$ marking the boundaries. The controllable sequence is constructed as:

$$seqE = \textcircled{EM}, \{E^l\}, \textcircled{/EM} \tag{3}$$

We train an autoregressive (AR) speech language model conditioned on the speaker embedding V, contextual $seqC$, controllable $seqE$, text tokens W^i, and GT tokens U^j, to predict the shifted speech token sequence $\tilde{U}_j$, formulated as:

$$\prod_{j=1}^{N} p(\tilde{U}_j \mid V, seqC, seqE, W, \tilde{U}_{<j}; \theta_{\text{AR}}) \tag{4}$$

We detail the training and inference strategies of our system below.

2.1. Contextual Consistency

While pre-context C_{pre} and post-context C_{post} intuitively enhance contextual consistency, training solely on textual context without explicit speaker modeling still requires prompt speech for voice specification during inference. Consequently, synthesized speech prosody remains constrained by the prompt speech, limiting semantic-prosodic alignment and narrative coherence in long-form synthesis.

Semantic Consistency Enhancement: To improve alignment between synthesized audio and text semantics while

mitigating excessive prosodic interference from prompt audio, we adopt a decoupled training strategy. Rather than the traditional coupled "prompt-target" training model, we employ independent modeling that focuses on the adaptive relationship between target audio prosody and text semantics. This approach reduces cross-audio prosodic interference at the data level and enhances consistency between synthesized audio and current text semantics.

Persona Consistency Enhancement: To avoid acoustic timbre shifts caused by insufficient prompt matching while decoupling, we propose a multi-constraint prompt selection mechanism with three refinements: 1) prompts are selected within the same chapter with voiceprint similarity constraints [14] to reduce cross-chapter variance; 2) an optimal similarity threshold is experimentally determined to balance voice stability and prosodic diversity; 3) high-quality labeled multi-emotional speech data is incorporated to alleviate sample sparsity and ensure consistent persona expression across emotions.

2.2. Controllability

This section introduces our controllability enhancement strategy, focusing on training and inference mechanisms.

Controllability Enhancement: During training, fiction or persona-derived instructions are decomposed into discrete attribute labels. For example, "shouting angrily" decomposes into "very angry, low volume, slow speed" for an ill elderly person, and "very angry, high volume, fast speed" for a healthy person—facilitating explicit attribute–acoustic feature mapping learning. During inference, a third module converts user instructions into attribute combinations. This discretization mitigates linguistic ambiguity and enhances control precision.

Emotional Intensity Enhancement: To alleviate the scarcity of high-intensity emotional samples, we employ a self-distillation strategy consisting of three key steps: First, samples with varying intensity levels are synthesized using a pre-trained emotional TTS model; Second, the generated samples are filtered through PER, speaker similarity and pitch to guarantee quality; Third, the intensity distribution is balanced via targeted data augmentation.

3. EXPERIMENTAL SETUP

3.1. Experimental Setup

Datasets and Training: There were three stages in our training. First, 1 million hours audiobook data was used to finetune CosyVoice2 for audiobook domain adaptation. Second, a 100K hours context-aware dataset and 500 hours recordings labeled with instructions were employed to further finetune the model. It is worthy to note that each sentence in context-aware dataset was annotated with timestamp, speaker IDs and position information in the whole chapter. Third, to enhance the expressiveness and controllability of instructions, a data augmentation dataset of 5k hours was constructed based on the model after two-stage training. During the training, the model was optimized with AdamW Optimizer [15] by setting learning rate to 1e-5 for first two stages but 1e-6 for the last stage. 64 NVIDIA A800 GPUs are employed to train the model with batch size of 384 for 720K, 300K and 10K steps respectively for three stages.

To evaluate on audiobook scenarios, three test sets were constructed: a narrative test set *Test-NAR* consists of 100 paragraphs with each paragraph over 240 sentences. A dialogue test set *Test-DIA* composed of 570 dialogue sentences. 15 chapters varying from 2000 to 4000 sentences were used to construct a chapter test set *Test-CHAP* including both narration and dialogue parts.

Model Evaluation: For comprehensive evaluation of our system, we used two subjective assessment methods [16]: Single-sentence Mean Opinion Score (S-MOS) for dialogue samples, and Multi-sentence Mean Opinion Score (M-MOS) for narration and long-chapter samples. We randomly sampled 20 paragraphs (from *Test-NAR*), 60 sentences (from *Test-DIA*) and 10 chapters (from *Test-CHAP*) for subjective ratings, with 50 Chinese native speakers recruited to score both S-MOS and M-MOS. We further compared it with two baselines via ABX tests: single-sentence (S-ABX) for dialogue, and multi-sentence (M-ABX) for narration and chapters.

Besides subjective evaluations, we use objective metrics, PER [17] and SS [14], to guarantee the stability of the model. Our proposed model was evaluated and compared under the following three configurations: **Infer-ctx**: inference using only the additional context sequence. **Infer-inst**: inference using only the additional instruction sequence. **Infer-ctx&inst**: inference incorporating both context and instruction sequences.

3.2. Main Results

The results from Table 1 and Fig.2 show that the proposed system largely exceeds both baseline models for all test setups. By comparing the results between the narration and dialogue tests, the results of the dialogue tests are more advantageous than those of the narration tests for both MOS and ABX tests. It is also worthy to mention that the speech generation strategy of *Infer-Ctx&Inst* does benefit from combining context-aware narration generation and instructed dialogue generation. From Fig.2, *Ctx&Inst* achieves the most preferable result compared to just *Infer-Ctx* and *Infer-inst*, which is consistent with the results in Table 1.

3.3. Ablation Analysis

To validate key components of our framework, four ablation experiments were conducted. Experiments 1–3 focus on context awareness, using a 100 k-hour audiobook chapter dataset for training, Test-DIA for evaluation, and SS [14], PER [17]

Table 1. The S-MOS and M-MOS of different context models

Model	Narration M-MOS	Dialogue S-MOS	Chapter M-MOS
Baseline			
CosyVoice2	3.91±0.07	3.84±0.09	3.88±0.07
MOSS-TTSD	3.78±0.11	3.67±0.07	3.72±0.09
Proposed			
Infer-ctx	**4.06±0.08**	3.93±0.06	4.13±0.09
Infer-inst	/	3.96±0.08	/
Infer-ctx&inst	/	**4.11±0.06**	**4.25±0.11**

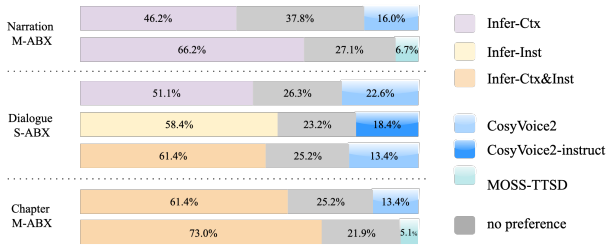


Fig. 2. Preference test results

and S-MOS as unified metrics. Experiment 4 focuses on instruction, adopting dedicated training data—including 500 hours of high-quality emotional data and 5,300 hours of augmented data—and tailored assessment methods.

Table 2. Performance Comparison of Context Strategy

Model	SS	PER	S-MOS
Non-Decoupled	**0.87**	1.33	3.45±0.09
Decoupled-0.68	0.69	1.19	3.86±0.06
Decoupled-0.8	0.79	1.84	3.82±0.07
Decoupled-0.8 + context	0.80	1.67	**3.93±0.06**

Decoupled vs Non-Decoupled Models: We compared two strategies for the context-aware module: the non-decoupled model, where audio prompts and targets are identical, and the decoupled model, in which distinct prompts and targets are selected via a specific strategy. Results in Table 2 show the non-decoupled model achieves excessively high SS [14], indicating over-similar timbre, prosody, and emotion—and thus impractical for audiobook due to insufficient character diversity.

Impact of Decoupling Threshold: Audio clips within chapters were clustered using different thresholds to generate distinct speaker IDs. A lower threshold reduced SS [14], with occasional timbre discontinuities but yielded slightly higher S-MOS than higher thresholds. Conversely, an excessively high threshold is hypothesized to increase SS [14], approaching non-decoupled model performance, while risking lower S-MOS, based on tested threshold trends.

Effect of Contextual Text Input: We evaluated the influence of contextual text on the context-aware module by feeding the Text-Speech Language Model with two input types: target text alone, or target text plus its preceding and subsequent sentence. The context-augmented model achieved a higher S-MOS, with listening tests confirming improved coherence. Next, we present an example.

> “叫师父给你多喂几碗符水，看你还胡说八道。我猜是师兄功力大进”
> While this sentence shows no explicit emotional indicators, the model naturally incorporated laughter into the synthesis based on the prior contextual “一旁的夏姝噗嗤笑个不停，”

Fine-Grained Emotional Control: We used the Chinese emotional speech test set from CV3-Eval [18] to evaluate controllability, modifying the original instructions into three types of emotional states: “high-intensity single emotion”, “low-intensity single emotion”, and “mixed emotion”. Three metrics are employed for emotional evaluation: emotion classification F1-score [19] for single emotions; S-MOS [16] and SS [16] for mixed emotions. Results in Table 3 show our model exhibits stronger H-L (discriminability between “high-intensity” and “low-intensity”) emotion control on “Text-Unrelated” test set and outperform baseline models [6] in high-intensity emotion control. Table 4 shows that our model also outperforms the baseline model [6] in terms of mixed emotion performance.

Table 3. Comparison of F1 scores For Single Emotion

Model	Text-Related			Text-Unrelated		
	angry	happy	sad	angry	happy	sad
CosyVoice2-H	0.79	0.96	0.68	0.07	0.62	0.53
CosyVoice2-L	0.77	0.92	0.68	0.00	0.68	0.38
Infer-inst-H	0.72	0.92	0.98	0.31	0.54	0.65
Infer-inst-L	0.62	0.91	0.94	0.00	0.39	0.32
$\Delta_{\text{CosyVoice2}}$(H-L)	0.02	0.04	0.00	0.07	-0.06	0.15
$\Delta_{\text{Infer-inst}}$(H-L)	**0.10**	0.01	**0.04**	**0.31**	**0.15**	**0.33**

Table 4. Performance Comparison For Mixed Emotion

Model	Text-Related		Text-Unrelated	
	SS	S-MOS	SS	S-MOS
CosyVoice2-instruct	0.74	3.67±0.06	0.75	3.35±0.07
Infer-inst	**0.77**	**4.08±0.07**	**0.78**	**3.87±0.09**

4. CONCLUSIONS

The paper proposes a controllable, context-aware TTS framework for multicast audiobooks, with 3 innovations: a context mechanism for contextual consistency, a disentanglement paradigm to decouple style control from speech prompts for semantic consistency, and self-distillation to boost emotional expressiveness and controllability. Experiments show the framework outperforms baselines in narration, dialogue, and chapter generation, with ablation studies validating its key components.In future, we can expand chapter data or select

specific data to mitigate data sparsity, and explore reinforcement learning for performance improvement.

5. ACKNOWLEDGEMENT

We acknowledge using Doubao [20] for this paper's language polishing, including grammar checks, context-appropriate term selection, and text flow improvement. Notably, Doubao was only used for language enhancements. All theoretical frameworks, empirical data work, and key arguments are the authors' independent efforts; we critically evaluated and integrated Doubao-generated suggestions to maintain academic authenticity and originality.

6. REFERENCES

[1] Kyeongman Park, Seongho Joo, and Kyomin Jung. Multiactor-audiobook: Zero-shot audiobook generation with faces and voices of multiple speakers. *arXiv preprint arXiv:2505.13082*, 2025.

[2] Yuxin Guo, Teng Wang, Yuying Ge, Shijie Ma, Yixiao Ge, Wei Zou, and Ying Shan. Audiostory: Generating long-form narrative audio with large language models. *arXiv preprint arXiv:2508.20088*, 2025.

[3] Yan Rong, Shan Yang, Guangzhi Lei, and Li Liu. Dopamine audiobook: A training-free mllm agent for emotional and human-like audiobook generation. *arXiv preprint arXiv:2504.11002*, 2025.

[4] Xuenan Xu, Jiahao Mei, Chenliang Li, Yuning Wu, Ming Yan, Shaopeng Lai, Ji Zhang, and Mengyue Wu. Mm-storyagent: Immersive narrated storybook video generation with a multi-agent paradigm across text, image and audio. *arXiv preprint arXiv:2503.05242*, 2025.

[5] Shaja Arul Selvamani and Nia D'Souza Ganapathy. A multi-agent ai framework for immersive audiobook production through spatial audio and neural narration, 2025.

[6] Zhihao Du, Yuxuan Wang, Qian Chen, Xian Shi, Xiang Lv, Tianyu Zhao, Zhifu Gao, Yexin Yang, Changfeng Gao, Hui Wang, et al. Cosyvoice 2: Scalable streaming speech synthesis with large language models. *arXiv preprint arXiv:2412.10117*, 2024.

[7] Zeqian Ju, Dongchao Yang, Jianwei Yu, Kai Shen, Yichong Leng, Zhengtao Wang, Xu Tan, Xinyu Zhou, Tao Qin, and Xiangyang Li. Mooncast: High-quality zero-shot podcast generation. *arXiv preprint arXiv:2503.14345*, 2025.

[8] OpenMOSS Team. Moss-ttsd: Text to spoken dialogue generation. *https://www.open-moss.com/en/moss-ttsd/*, 2025.

[9] Leying Zhang, Yao Qian, Long Zhou, Shujie Liu, Dongmei Wang, Xiaofei Wang, Midia Yousefi, Yanmin Qian, Jinyu Li, Lei He, et al. Covomix: Advancing zero-shot speech generation for human-like multi-talker conversations. *Advances in Neural Information Processing Systems*, 37:100291–100317, 2024.

[10] Shehzeen Hussain, Paarth Neekhara, Xuesong Yang, Edresson Casanova, Subhankar Ghosh, Mikyas T Desta, Roy Fejgin, Rafael Valle, and Jason Li. Koel-tts: Enhancing llm based speech generation with preference alignment and classifier free guidance. *arXiv preprint arXiv:2502.05236*, 2025.

[11] Charuta Pethe, Bach Pham, Felix D Childress, Yunting Yin, and Steven Skiena. Prosody analysis of audiobooks. *arXiv preprint arXiv:2310.06930*, 2023.

[12] Dake Guo, Xinfa Zhu, Liumeng Xue, Yongmao Zhang, Wenjie Tian, and Lei Xie. Text-aware and context-aware expressive audiobook speech synthesis. *arXiv preprint arXiv:2406.05672*, 2024.

[13] Zhipeng Li, Xiaofen Xing, Jingyuan Xing, Hangrui Hu, Heng Lu, and Xiangmin Xu. Long-context speech synthesis with context-aware memory. *arXiv preprint arXiv:2508.14713*, 2025.

[14] Hui Wang, Siqi Zheng, Yafeng Chen, Luyao Cheng, and Qian Chen. Cam++: A fast and efficient network for speaker verification using context-aware masking. *arXiv preprint arXiv:2303.00332*, 2023.

[15] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

[16] Shun Lei, Yixuan Zhou, Liyang Chen, Zhiyong Wu, Shiyin Kang, and Helen Meng. Context-aware coherent speaking style prediction with hierarchical transformers for audiobook speech synthesis, 2023.

[17] Smith M and et al. Automating error frequency analysis via the phonemic edit distance ratio. *J Speech Lang Hear Res. 2019 Jun 19;62(6):1719-1723.*, 2019.

[18] Zhihao Du, Changfeng Gao, Yuxuan Wang, Fan Yu, Tianyu Zhao, Hao Wang, Xiang Lv, Hui Wang, Xian Shi, Keyu An, et al. Cosyvoice 3: Towards in-the-wild speech generation via scaling-up and post-training. *arXiv preprint arXiv:2505.17589*, 2025.

[19] Ziyang Ma and et al. emotion2vec: Self-supervised pre-training for speech emotion representation. *arXiv preprint arXiv:2312.15185*, 2023.

[20] Ziheng Jiang and et al. Megascale: Scaling large language model training to more than 10,000 gpus. *arXiv preprint arXiv:2402.15627*, 2024.