Proceedings of the Thirty-Fifth International Joint Conference on Artificial Intelligence (IJCAI-26)
Demonstrations Track

Low-Latency Real-Time Audio Game Commentary System via LLM-Based Parallel Text Generation

Ryota Kawamatsu[1,2], **Anum Afzal**[2,3], **Yuki Saito**[1], **Shinnosuke Takamichi**[4,1], **Graham Neubig**[5], **Katsuhito Sudoh**[6], **Hiroya Takamura**[2], **Tatsuya Ishigaki**[2]
[1]The University of Tokyo, Japan
[2]National Institute of Advanced Industrial Science and Technology, Japan
[3]Technical University of Munich, Germany
[4]Keio University, Japan
[5]Carnegie Mellon University, U.S.A.
[6]Nara Women's University, Japan

Abstract

We present a low-latency real-time audio game commentary system that generates spoken commentary directly from live gameplay video. In this end-to-end setting, a key bottleneck is accumulated waiting time; conventional pipelines capture frames, generate text, and synthesize speech sequentially for each utterance, and do not request the next generation until speech playback has completed. This strict sequentiality causes long and unnatural silence between utterances. To address this latency bottleneck, our system runs text generation in parallel with speech playback and buffers multiple candidate utterances ahead of time, enabling immediate synthesis at playback boundaries. Experiments on fast-paced game videos show that our parallel design reduces the mean inter-utterance silence from 9.6 seconds to 0.3 seconds compared to sequential baselines. It also improves similarity to professional speaking–silence timing patterns by over 40 %, and a user study with 120 experienced game players confirms significantly improved perceived speaking rhythm. Our demo video is available at: https://youtu.be/pmrRUlvav8M.

1 Introduction

Live game commentary describes in-game events and provides contextual explanations to help viewers immerse themselves in gameplay [Behrens and Uhrich, 2022; da Silva and Scelles, 2025]. Audio commentary is integrated with the live-streamed video and delivered to users via online streaming platforms. Advances in multimodal language models and audio synthesis have accelerated research on automatic audio commentary generation [Zheng *et al.*, 2025]. Prior work has explored language generation from structured data [Taniguchi *et al.*, 2019; Ishigaki *et al.*, 2021], video [Yu *et al.*, 2018; Kim and Choi, 2020; Rao *et al.*, 2024; Zhou *et al.*, 2024; Chen *et al.*, 2025; Afzal *et al.*, 2026], and multimodal inputs [Mori *et al.*, 2025; Someya *et al.*, 2025; Wang and Yoshinaga, 2025], as well as expressive speech synthesis for engaging commentary[Iura *et al.*, 2025]; however, these two lines of research have mostly progressed independently.

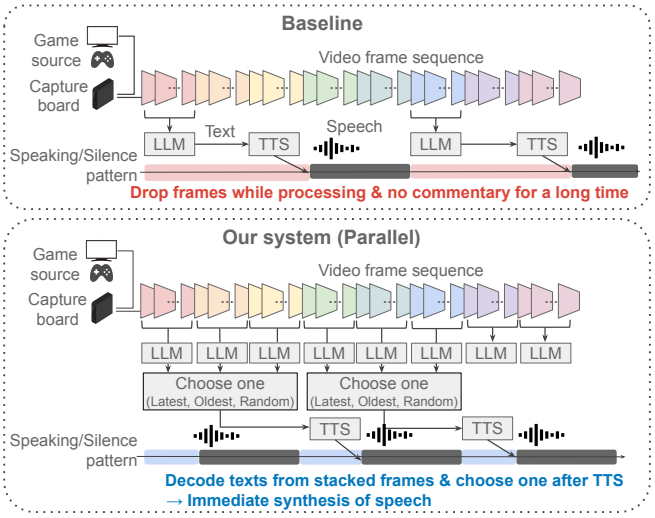

Figure 1: System overview: sequential baseline vs. our parallel generation with buffering and video delay control.

Building on these separate research directions, several works have begun to combine language generation and speech synthesis into end-to-end audio commentary systems [Kumano *et al.*, 2019; Jumneanbun *et al.*, 2020; Xu *et al.*, 2021; Ishigaki *et al.*, 2023]. However, existing systems rely on structured game state inputs and lightweight language models, which prevent latency from becoming a critical bottleneck. Extending this setting to raw gameplay video and modern multimodal large language models (LLMs) introduces substantial latency challenges [Mori *et al.*, 2025].

A major bottleneck in real-time audio commentary is its pipeline design. Most pipelines are sequential: they process game inputs, generate text, synthesize speech, and start generating the next utterance only after the current speech playback has completed. As a result, waiting time accumulates across turns and leads to long silence between utterances, which sub-

stantially degrades perceived commentary quality.

To address this latency bottleneck, we propose an audio commentary system with parallel text generation and a lightweight control of video delay. While one utterance is being spoken, our system continues generating candidate utterances in the background process for newly arriving video segments and stores them in a buffer. When the current speech playback ends, the system immediately selects a buffered candidate and synthesizes speech from it, thereby avoiding silence caused by waiting for generation to finish. Since inference latency is inevitable in practice, we additionally implement a lightweight video delay control that intentionally delays the outgoing video stream to better align generated speech with the displayed visuals.

Experiments on fast-paced game videos [Saito *et al.*, 2020] show that our system reduces mean inter-utterance silence from 9.6 seconds to 0.3 seconds, a gain further supported by a human evaluation with 120 participants.

2 Real-Time Game Audio Commentary

A typical real-time audio commentary pipeline converts a gameplay stream into audio commentary by executing (i) video acquisition and segmentation, (ii) text generation, and (iii) speech synthesis. The video stream is buffered and grouped into short frame segments. Let f_i denote the frame captured at time i; then a segment is $F_k = \{f_i, \ldots, f_{i+N-1}\}$, which consists of N consecutive frames. A common design is to run these modules sequentially. Specifically, the system waits until the current speech playback ends, and then calls a multimodal LLM on the latest segment to generate text using a simple prompt (e.g., "Describe the game situation. If you have nothing to say, stay silent.") [Afzal *et al.*, 2026]. The generated text is then converted into audio by a text-to-speech (TTS) module [Iura *et al.*, 2025]. Because the next generation is triggered only after playback finishes, this strict sequentiality accumulates latency and causes long silences between utterances.

3 Proposed System

We propose a low-latency system design based on two key ideas: parallel text generation (our main mechanism to reduce waiting time) and lightweight video delay control (to maintain temporal consistency).

3.1 Parallel Text Generation

Unlike conventional systems that trigger text generation only after speech playback ends, our system initiates text generation as soon as a new video segment becomes available, even while the current utterance is still being spoken (Figure 1). As a result, multiple candidate utterances are generated ahead of playback and stored in a buffer. When the current utterance ends, the system selects one buffered candidate and synthesizes it immediately, eliminating idle silence caused by sequential waiting.

We consider three lightweight selection policies: *Latest* (most recent segment), *Oldest* (earliest buffered), and *Random*. We adopt these lightweight policies to avoid additional decision latency in real-time settings.

3.2 Video Delay Control

We consider a streaming setup for live streaming platforms such as YouTube Live, where gameplay video and automatically generated speech are integrated into a single audiovisual stream on our server and then uploaded to the platform. This setting allows us to control video playback timing to mitigate misalignment caused by generation latency.

Since text generation and TTS inevitably introduce delays that cause commentary to lag, we absorb this latency by intentionally delaying the video. Specifically, we buffer the video stream and start playback only when the first utterance begins, matching the initial end-to-end generation latency.

4 Experiments

Demo Setups: In the demonstration, conference attendees voluntarily operate a Nintendo Switch console on-site, and the gameplay video is captured in real time using a capture device (Elgato HD60 X[1]) and streamed to a commentary server. The system generates spoken commentary online based on the live video input. The demo is conducted exclusively on-site, and no gameplay video or audio is recorded or distributed. Attendees observe the gameplay together with automatically generated commentary, allowing them to directly perceive how parallel candidate generation affects inter-utterance silence and speaking rhythm. The gameplay machine sends video at 25 fps, and captured frames are grouped into segments of N consecutive frames. In all experiments, we set $N = 32$. Each segment $F_k = \{f_i, \ldots, f_{i+N-1}\}$ is sent to the commentary server as soon as it becomes available. The language model receives each segment as base64-encoded tokens and generates commentary text. We use *GPT-4.1-mini* [2] for text generation and follow the TTS configuration described in [Iura *et al.*, 2025]. Our experiments vary the `max_new_tokens` parameter in $\{20, 40, 60, 80, 100\}$.

Dataset: We randomly selected 8 videos from the Smash Corpus [Saito *et al.*, 2020] whose target game is *Super Smash Bros. Ultimate*. We select this game for our experiments because it is a fast-paced game where latency is particularly noticeable. This corpus provides gameplay video. We collected commentaries from skilled commentators for the selected videos, enabling a comparison between skilled human- and system-generated commentaries.

Compared Models: We compare five methods. All methods use the same video delay control (Section 3.2). *Baseline After-Audio* is a fully sequential pipeline that starts generating the next utterance only after the current speech playback ends, following a common design in prior real-time commentary systems [Ishigaki *et al.*, 2023]. *Baseline After-Text* is a semi-sequential variant that starts generating the next utterance immediately after text generation finishes, without waiting for speech playback to end. Our system uses parallel generation with buffering, and we test three lightweight selection policies: *Parallel Latest*, *Prallel Oldest*, and *Parallel Random*.

[1] https://www.elgato.com/jp/ja/p/game-capture-hd60-x

[2] https://platform.openai.com/docs/models/gpt-4.1-mini

Method	Cumulative silence	Mean silence	mIoU
Human	59.7 ± 11.3	1.7 ± 0.4	–
After-Audio	134.0 ± 5.2	9.5 ± 0.9	0.01 ± 0.06
After-Text	125.5 ± 5.8	6.8 ± 0.9	0.10 ± 0.08
Latest	24.7 ± 6.6	0.4 ± 0.1	0.59 ± 0.04
Oldest	$\mathbf{18.9} \pm 3.1$	$\mathbf{0.3} \pm 0.1$	$\mathbf{0.60} \pm 0.04$
Random	19.1 ± 3.6	$\mathbf{0.3} \pm 0.1$	$\mathbf{0.60} \pm 0.03$

Table 1: Silence statistics and mIoU for `max_new_tokens`=20. Values in parentheses denote standard deviation.

`max_new_tokens`	Mean speaking time	Text length
Human	2.8 ± 0.5	25.9 ± 4.1
20	2.3 ± 0.1	16.8 ± 0.4
40	2.9 ± 0.1	22.7 ± 0.6
60	3.4 ± 0.1	28.2 ± 0.8
80	3.8 ± 0.1	31.8 ± 1.1
100	4.4 ± 0.1	36.3 ± 1.3

Table 2: Effect of `max_new_tokens` on speaking duration and text length (Latest policy).

Evaluation Metrics: We evaluate commentaries in terms of timing behavior and content adequacy. For timing, we compare cumulative silence, mean silence between utterances, and utterance length. To capture similarity of speaking and silence patterns to human-spoken commentaries, we represent each commentary stream as a 1 Hz binary sequence (1 = speaking, 0 = silence) and compute the mean Intersection over Union (mIoU) between automatically generated commentaries and skilled human-spoken ones. Higher mIoU indicates fewer mismatches in speaking and silence regions compared to skilled human-spoken commentaries. For content adequacy, we compute ROUGE [Lin, 2004] scores against reference commentary text. Because exact temporal alignment is difficult in real-time settings, ROUGE is computed within fixed 10-second windows.

For human evaluation, we conduct a user study with 120 crowdworkers recruited via Lancers[3]. We present 30-second clips sampled from 20 scenes. Each clip is paired with commentary generated by one of 25 conditions (5 methods $\times$ 5 `max_new_tokens` settings). Participants rate on 5-point Likert scales, and we report Mean Opinion Scores (MOS) for the following criteria: (Q1) commentary rhythm naturalness, (Q2) alignment with the video, and (Q3) overall quality.

5 Results

Table 1 summarizes silence-related statistics and mIoU for each method. The sequential baselines (After-Audio and After-Text) exhibit approximately twice the cumulative silence of human commentary and extremely low mIoU scores, indicating that sequential processing introduces long and unnatural pauses between utterances. In contrast, all proposed parallel methods substantially reduce silence, achieving mIoU values close to human commentary. Differences among the selection policies are relatively small, suggesting that the main benefit comes from parallel generation itself.

[3]https://www.lancers.jp/

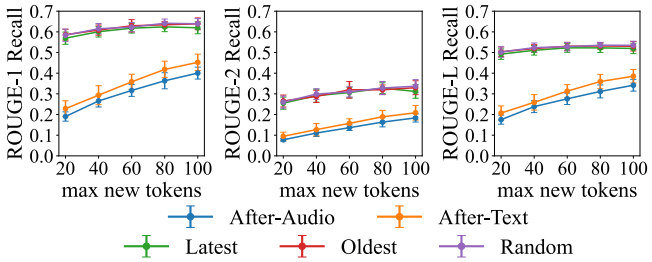

Figure 2: ROUGE Recall across `max_new_tokens`. Error bars denote standard deviation.

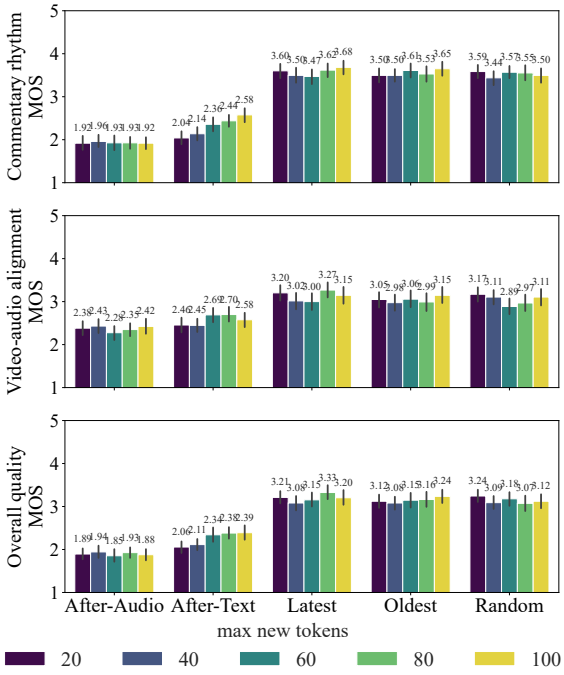

Figure 3: User study results for Q1–Q3 across methods and `max_new_tokens`. Error bars denote 95% confidence intervals.

Table 2 analyzes the effect of `max_new_tokens` under the Latest policy. While a setting of 40 best matches human speaking duration, 60 yields utterance lengths closest to human commentary, indicating that generation length affects different aspects of temporal behavior. ROUGE scores (Figure 2) show consistent improvements of parallel methods over sequential baselines.

Figure 3 reports the results of the human evaluation. The proposed methods significantly outperform the baselines across all criteria (Q1–Q3), demonstrating that reducing unnatural silence leads to higher perceived commentary quality.

6 Conclusion

We presented a real-time game audio commentary system that reduces latency by parallel generation. Experiments show substantial reductions in inter-utterance silence and improved similarity to professional speaking/silence patterns.

Acknowledgements

This paper is based on results obtained from a project, Programs for Bridging the gap between R&D and the IDeal society (society 5.0) and Generating Economic and social value (BRIDGE)/Practical Global Research in the AI × Robotics Services, implemented by the Cabinet Office, Government of Japan. This paper is also based on results obtained from AIST policy-based budget project "R&D on Generative AI Foundation Models for the Physical Domain".

References

[Afzal *et al.*, 2026] Anum Afzal, Yuki Saito, Hiroya Takamura, Katsuhito Sudoh, Shinnosuke Takamichi, Graham Neubig, Florian Matthes, and Tatsuya Ishigaki. Real-time generation of game video commentary with multimodal llms: Pause-aware decoding approaches. In *Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026)*, pages 9188–9201, May 2026.

[Behrens and Uhrich, 2022] Anton Behrens and Sebastian Uhrich. You'll never want to watch alone: the effect of displaying in-stadium social atmospherics on media consumers' responses to new sport leagues across different types of media. *European Sport Management Quarterly*, 22(1):120–138, 2022.

[Chen *et al.*, 2025] Joya Chen, Ziyun Zeng, Yiqi Lin, Wei Li, Zejun Ma, and Mike Zheng Shou. LiveCC: Learning video llm with streaming speech transcription at scale. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 29083–29095, June 2025.

[da Silva and Scelles, 2025] Jaime A. Teixeira da Silva and Nicolas Scelles. A vision for the formal documentation and digitalization of sports commentators' commentaries. *International Journal of Sport Communication*, 18(1):1–9, 2025.

[Ishigaki *et al.*, 2021] Tatsuya Ishigaki, Goran Topic, Yumi Hamazono, Hiroshi Noji, Ichiro Kobayashi, Yusuke Miyao, and Hiroya Takamura. Generating racing game commentary from vision, language, and structured data. In *Proceedings of the 14th International Conference on Natural Language Generation*, pages 103–113, August 2021.

[Ishigaki *et al.*, 2023] Tatsuya Ishigaki, Goran Topic, Yumi Hamazono, Ichiro Kobayashi, Yusuke Miyao, and Hiroya Takamura. Audio commentary system for real-time racing game play. In *Proceedings of the 16th International Natural Language Generation Conference: System Demonstrations*, pages 9–10, September 2023.

[Iura *et al.*, 2025] Kota Iura, Yuki Saito, Shinnosuke Takamichi, Graham Neubig, Katsuhito Sudoh, Hiroshi Saruwatari, Hiroya Takamura, and Tatsuya Ishigaki. Excitement-inducing commentary text-to-speech system for fighting game video scenes. *IEEE Access*, 13:216748–216758, 2025.

[Jumneanbun *et al.*, 2020] Thanat Jumneanbun, Sunee Saelao, Pujana Paliyawan, Ruck Thawonmas, Kingkarn Sookhanaphibarn, and Worawat Choensawat. Rap-style comment generation to entertain game live streaming. In *Proceedings of the IEEE Conference on Games (CoG)*, pages 706–707, 2020.

[Kim and Choi, 2020] Byeong Jo Kim and Yong Suk Choi. Automatic baseball commentary generation using deep learning. In *Proceedings of the 35th Annual ACM Symposium on Applied Computing*, page 1056–1065, 2020.

[Kumano *et al.*, 2019] Tadashi Kumano, Manon Ichiki, Kiyoshi Kurihara, Hiroyuki Kaneko, Tomoyasu Komori, Toshihiro Shimizu, Nobumasa Seiyama, Atsushi Imai, Hideki Sumiyoshi, and Tohru Takagi. Generation of automated sports commentary from live sports data. In *2019 IEEE International Symposium on Broadband Multimedia Systems and Broadcasting (BMSB)*, pages 1–4, 2019.

[Lin, 2004] Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, July 2004.

[Mori *et al.*, 2025] Yuichiro Mori, Chikara Tanaka, Aru Maekawa, Satoshi Kosugi, Tatsuya Ishigaki, Kotaro Funakoshi, Hiroya Takamura, and Manabu Okumura. Live football commentary system providing background information. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 394–404, July 2025.

[Rao *et al.*, 2024] Jiayuan Rao, Haoning Wu, Chang Liu, Yanfeng Wang, and Weidi Xie. MatchTime: Towards automatic soccer game commentary generation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 1671–1685, November 2024.

[Saito *et al.*, 2020] Yuki Saito, Shinnosuke Takamichi, and Hiroshi Saruwatari. SMASH corpus: A spontaneous speech corpus recording third-person audio commentaries on gameplay. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 6571–6577, May 2020.

[Someya *et al.*, 2025] Taiga Someya, Tatsuya Ishigaki, and Hiroya Takamura. Live football commentary (LFC): A Large-Scale dataset for building football commentary generation models. In *Proceedings of the 18th International Natural Language Generation Conference*, pages 195–201, October 2025.

[Taniguchi *et al.*, 2019] Yasufumi Taniguchi, Yukun Feng, Hiroya Takamura, and Manabu Okumura. Generating live soccer-match commentary from play data. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2019.

[Wang and Yoshinaga, 2025] Zihan Wang and Naoki Yoshinaga. Commentary generation from multimodal game data for esports moments in multiplayer strategy games. In *Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, pages 1795–1807, December 2025.

[Xu *et al.*, 2021] Junjie H. Xu, Zhou Fang, Qihang Chen, Satoru Ohno, and Pujana Paliyawan. Fighting game commentator with pitch and loudness adjustment utilizing

highlight cues. In *Proceedings of the IEEE Global Conference on Consumer Electronics (GCCE)*, pages 366–370, 2021.

[Yu *et al.*, 2018] Huanyu Yu, Shuo Cheng, Bingbing Ni, Minsi Wang, Jian Zhang, and Xiaokang Yang. Fine-grained video captioning for sports narrative. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6006–6015, 2018.

[Zheng *et al.*, 2025] Qirui Zheng, Xingbo Wang, Keyuan Cheng, Muhammad Asif Ali, Yunlong Lu, and Wenxin Li. From multimodal perception to strategic reasoning: A survey on ai-generated game commentary, 2025.

[Zhou *et al.*, 2024] Xingyi Zhou, Anurag Arnab, Shyamal Buch, Shen Yan, Austin Myers, Xuehan Xiong, Arsha Nagrani, and Cordelia Schmid. Streaming dense video captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18243–18252, June 2024.