

Mid-Utterance Transition Control for Audio Commentary Generation

Ryota Kawamatsu^{1,2}, Tatsuya Ishigaki², Hiroya Takamura²

¹The University of Tokyo

²National Institute of Advanced Industrial Science and Technology

{kawamatsu.ryota, takamura.hiroya}@aist.go.jp

ishigaki_tatsuya@cyberagent.co.jp

Abstract

Audio commentary generation enhances spectator engagement by describing ongoing events. A critical challenge in this domain is promptly reacting to newly emerging events while maintaining coherent transitions. Existing approaches face a fundamental trade-off: waiting for utterance completion causes delays, whereas immediate transitions risk unnatural mid-word truncation. Despite its importance, mid-utterance transition control remains under-explored. We formulate this problem as deciding when to terminate an ongoing utterance and how to connect it to the next. We compare four human-inspired strategies: immediate interruption, chunk-based interruption at syntactic boundaries, LLM-inserted break points, and LLM-generated shortened closures. Our experiments on a fast-paced game quantify the trade-off between responsiveness and naturalness. Human evaluation reveals that natural transitions do not always require complete sentences; interrupting at clear linguistic boundaries preserves responsiveness while avoiding unnatural truncation.

1 Introduction

Audio commentary provides spoken descriptions of ongoing game events for spectators. Beyond conveying factual information, effective commentary enhances immersion and entertainment through timely reactions, expressive delivery, and domain-specific interpretation (Behrens and Uhrich, 2022; da Silva and Scelles, 2025). These requirements have motivated growing interest in automated commentary generation using language generation and multimodal understanding technologies (Zheng et al., 2025).

Most existing systems follow a sequential pipeline of language generation and text-to-speech. Given the current video context, a vision-language model or large language model generates a textual utterance, which is then synthesized and played

Existing pipeline method

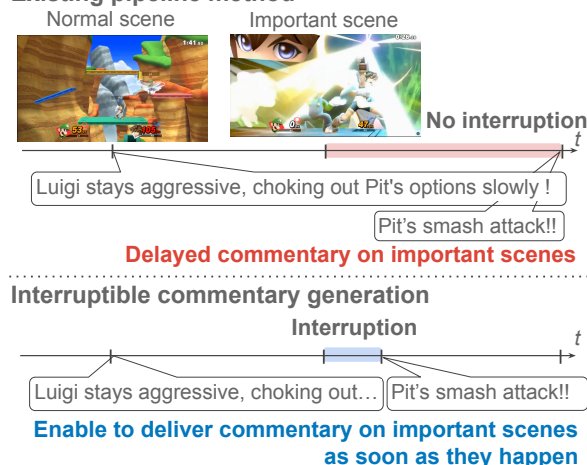


Figure 1: Overview of the interruption problem in live commentary generation. Without interruption, commentary on an important event is delayed until the ongoing utterance finishes. Interrupting the current utterance enables the system to react immediately, but introduces the challenge of transitioning naturally between the interrupted and newly generated speech.

using text-to-speech (Kawamatsu et al., 2026). Because the system plays each utterance to completion, this design preserves linguistic completeness and produces natural individual utterances. However, once playback begins, the system typically cannot react to an emerging event until the current utterance finishes, causing decisive plays to go unmentioned or be described with substantial delay.

A simple approach to improve responsiveness is to interrupt the ongoing utterance and immediately begin describing the new event. Some existing commentary systems have incorporated such interruption mechanisms (Tanaka-Ishii et al., 1998; Ishigaki et al., 2023). However, this approach sacrifices linguistic naturalness for responsiveness, as illustrated in Figure 1. Waiting for the ongoing utterance to finish preserves grammatical and acoustic

continuity but delays the response. At the other extreme, stopping playback immediately minimizes latency but may truncate speech in the middle of a word or syntactic unit, producing an abrupt and incomplete transition. The system must therefore determine not only *when* a new event should trigger commentary, but also *how* the ongoing utterance should transition to the new one.

Despite its practical importance, this transition problem has received limited attention. It remains unclear whether a system should stop immediately, wait for an appropriate boundary, or revise the unfinished utterance. These alternatives likely balance responsiveness and naturalness differently, yet they have not been systematically compared.

In this paper, we study *mid-utterance transition control* for live audio commentary generation. Motivated by interruption patterns in human commentary, we compare four transition strategies: 1) immediate interruption at the event trigger, 2) chunk-based interruption at syntactic chunk boundaries, 3) interruption at LLM-inserted <break> points, and 4) LLM-generated shortened closures replacing the unspoken remainder. We further evaluate each strategy with and without an interjection preceding the new commentary.

We compare these strategies using 40 interruption instances constructed from eight videos of a fast-paced fighting game. Both automatic and human evaluations confirm a clear trade-off between responsiveness and naturalness among the methods. However, human feedback indicates that linguistic completeness is not strictly necessary for a natural transition.

Our contributions are as follows:

- We formulate mid-utterance transition as a problem in live audio commentary generation.
- Motivated by interruption patterns in human commentary, we implement and compare four transition strategies that differ in interruption timing and linguistic completion.
- We find a clear tradeoff between responsiveness and linguistic naturalness.
- We show that linguistic completeness is not always necessary for a natural transition: stopping at a recognizable chunk boundary, particularly with an interjection, provided the most effective balance among the evaluated strategies in our current pipeline.

2 Related Work

Commentary generation has been studied across a wide variety of domains, including soccer (Sun et al., 2026; Someya et al., 2025; Mori et al., 2025), baseball (Kim and Choi, 2020), motorsports (Ishigaki et al., 2021, 2023; Afzal et al., 2026), and e-sports (Kawamatsu et al., 2026; Afzal et al., 2026). Existing work also differs in the input modality, including video (Sun et al., 2026; Afzal et al., 2026), play-by-play descriptions (Wang and Yoshinaga, 2025; Taniguchi et al., 2019), and numerical time series (Ishigaki et al., 2021; Someya et al., 2025).

Early work of commentary generation relied on chunk-based or template-driven methods that triggered predefined utterances under manually designed conditions (Tanaka-Ishii et al., 1998). As machine learning-based approaches, neural network-based encoder-decoder methods enabled more flexible generation by learning mappings from multimodal inputs to commentary (Taniguchi et al., 2019; Ishigaki et al., 2021). More recently, in-context learning with vision-language models has become a common paradigm (Yu et al., 2018; Kim and Choi, 2020; Rao et al., 2024; Zhou et al., 2024; Chen et al., 2025; Wang and Yoshinaga, 2025; Afzal et al., 2026; Sun et al., 2026; Kawamatsu et al., 2026). We focus on LLM-based video-to-commentary generation for fast-paced games, where maintaining natural transitions between utterances are expected to be particularly important and challenging.

Utterance transition in automated commentary generation has received limited attention. Most existing video-to-commentary systems wait until playback finishes before generating the next utterance (Afzal et al., 2026; Kawamatsu et al., 2026). Some exceptions include the reactive commentary systems of Tanaka-Ishii et al. (Tanaka-Ishii et al., 1998) and Ishigaki et al. (Ishigaki et al., 2023), which detect important events and trigger predefined utterances. While these approaches study “when” an interruption should occur, they do not investigate “how” the ongoing utterance should transition to a newly generated utterance. Our work addresses this gap through a comparative human evaluation of multiple transition strategies.

Conceptually, adapting an ongoing utterance to reflect new information relates to the phenomenon of *self-repair*, explored in early incremental natural language generation (Hough, 2011). In such systems, self-repair typically involves dynamically

employing conjunctions (e.g., “however” or “but”) to grammatically and semantically merge an ongoing context with a newly occurred, sometimes conflicting event into a single continuous sentence. In contrast, our current work specifically targets utterance *interruption*. Rather than attempting full grammatical repair, we investigate how to naturally break the ongoing speech and seamlessly transition to a newly generated utterance using strategies such as interjections and break tags.

Apart from language generation studies, spoken dialogue systems have long employed “barge-in” mechanisms that terminate agent output once a user begins speaking (Heins et al., 1997). As in commentary generation, *how* the output is terminated has received far less attention than *when* (Skantze, 2021). Early systems cut playback immediately at the detected onset (Ström and Seneff, 2000), and incremental dialogue managers were developed to dynamically halt and revise output in response to such user interruptions (Buß and Schlangen, 2011). Subsequent approaches re-synthesize the ongoing phrase to splice it at a syllable boundary (Wester et al., 2017), while recent full-duplex models predict interruption timing in a data-driven manner (Défossez et al., 2024; Veluri et al., 2024). These alternatives have not been compared under a common evaluation. Our setting also differs in its trigger: dialogue interruption is initiated by human speech and the system yields the floor by falling silent, whereas commentary requires *self-interruption* driven by the system’s own observation of the game, and is immediately followed by new speech.

3 Problem Formulation

We first describe the ideal real-time commentary generation setting considered in this work and then formulate the mid-utterance interruption problem.

3.1 Live Commentary Generation System

The conceptual real-time commentary generation system operates in two concurrent processes. First, the system analyzes the input game video at one-second intervals. For each interval, it predicts a single scene-importance score and generates the three types of candidate commentary utterances described in Section 3.2. Second, the system selects and plays audio of commentary utterances as the video progresses. A new utterance begins under one of two conditions: (1) the playback of the cur-

rent utterance has finished, or (2) a newly analyzed interval is assigned the highest importance score while another utterance is still being played.

Under condition (1), the system randomly selects one of the three candidates associated with the current video interval and begins playing it.¹ This produces a non-interruptive transition. Under condition (2), the system transitions from the ongoing utterance to commentary generated for the important scene, requiring a mid-utterance interruption. This paper focuses on how to realize the transition under condition (2).

3.2 Commentary Utterance Generation

The input game video is processed at one-second intervals. For each target timestamp, the system uses the 30 frames from the preceding one-second window, corresponding to a frame rate of 30 fps. Let

$$F_k = (f_{i-29}, \dots, f_i)$$

denote the k -th frame sequence, where f_i is the frame at the target timestamp and F_k contains the 30 consecutive frames covering the preceding second. For each frame sequence F_k , a multimodal large language model generates three types of commentary candidates that serve different roles in live commentary:

- **Reaction:** a short, spontaneous utterance expressing surprise or excitement about a newly occurring event.
- **Play-by-play Commentary:** a concise description of the ongoing player actions and game situation.
- **Color Commentary:** analytical or explanatory commentary providing tactical insight or background information.

The prompts we used are shown in Appendix A.

Along with these commentary candidates, the model predicts an importance score for the scene:

$$s_k \in \{0, 1, 2\},$$

where 0, 1, and 2 represent normal, slightly important, and decisive important events, respectively.

¹Utterance selection is itself an important problem in commentary generation. Because this paper focuses on utterance transition rather than utterance selection, we use the simplest strategy of random selection. More sophisticated approaches, such as diversity-based selection (Ishigaki et al., 2023), could also be incorporated.

#	Japanese Commentary	English Translation
1	ただ突進力がありますからそれをうまく使ってますねダメージ稼いで、あっと、ミュウツーが	He does have strong rushing power, so he can use it to rack up damage... ah , Mewtwo is—!
2	そうこうしているうちにでもゼルダもダメージが溜まってきました何とかあー爆弾に当ててしまった自ら	Zelda’s damage is piling up too. She’s gotta... oh no! She hit the bomb herself!

Table 1: Examples of mid-utterance transitions in the SMASH corpus. Bold text indicates interjections used to signal an unexpected event and transition to the subsequent commentary.

Among the generated candidates, one commentary utterance is selected and synthesized by a text-to-speech model for audio playback. Let this utterance be denoted by C_k .

While C_k is being synthesized and spoken, the system continues processing subsequent frame sequences in the background. For each newly arriving frame sequence, the LLM generates updated commentary candidates and predicts their importance scores without waiting for the current utterance to finish. Consequently, a newly detected high-priority event may arise while another utterance is still being played, requiring the system to decide how to interrupt the ongoing utterance.

3.3 Mid-Utterance Interruption

Suppose that commentary utterance C_k is currently being played by the text-to-speech model. During its playback, the system continues analyzing subsequent frame sequences and generating new commentary candidates. If a later frame sequence F_{k+m} is assigned the highest importance score 2, the system selects the newly generated commentary utterance C_{k+m} for playback. The challenge is the playback of C_k may still be in progress when C_{k+m} becomes available.

The task of mid-utterance interruption is to determine

1. when the playback of the ongoing utterance C_k should terminate, and
2. how the system should transition from C_k to the newly generated utterance C_{k+m} .

4 Methods

This section presents the mid-utterance transition strategies investigated in this study. To better understand how such transitions occur in practice, we first examine how human commentators adapt an ongoing utterance when an important event suddenly emerges. Based on the observed behaviors,

we implement four strategies that differ in where they terminate or revise the ongoing utterance before transitioning to the newly selected commentary.

4.1 Interruptions in human commentary

We qualitatively examined mid-utterance interruptions in the SMASH corpus (Saito et al., 2020). Our analysis was motivated by a possible mismatch between human commentary and existing automated interruption mechanisms. Previous systems interrupt the ongoing audio as soon as an important event is detected (Tanaka-Ishii et al., 1998; Ishigaki et al., 2023). Because the stopping point is determined by event timing rather than linguistic structure, this strategy can truncate speech at an arbitrary acoustic position, including in the middle of a word.

In contrast, we observed that human commentators rarely terminate an utterance at an arbitrary position. When shifting attention to an unexpected event, they typically abandon the ongoing utterance at a word or phrase boundary. As illustrated in Table 1, the interruption is also often followed by an interjection such as “ah” or “oh,” which signals the sudden event and cushions the transition to the subsequent commentary.

This observation motivates two design choices examined in this study. First, an automated system should avoid arbitrary mid-word truncation and instead interrupt speech at a linguistically meaningful boundary. Second, an interjection may help reduce the perceived abruptness of the transition. The following strategies implement these choices using different mechanisms for determining the interruption point.

4.2 Transition Strategies

We investigate the following four strategies.

Immediate interruption: The system stops the currently playing audio as soon as C_{k+m} becomes

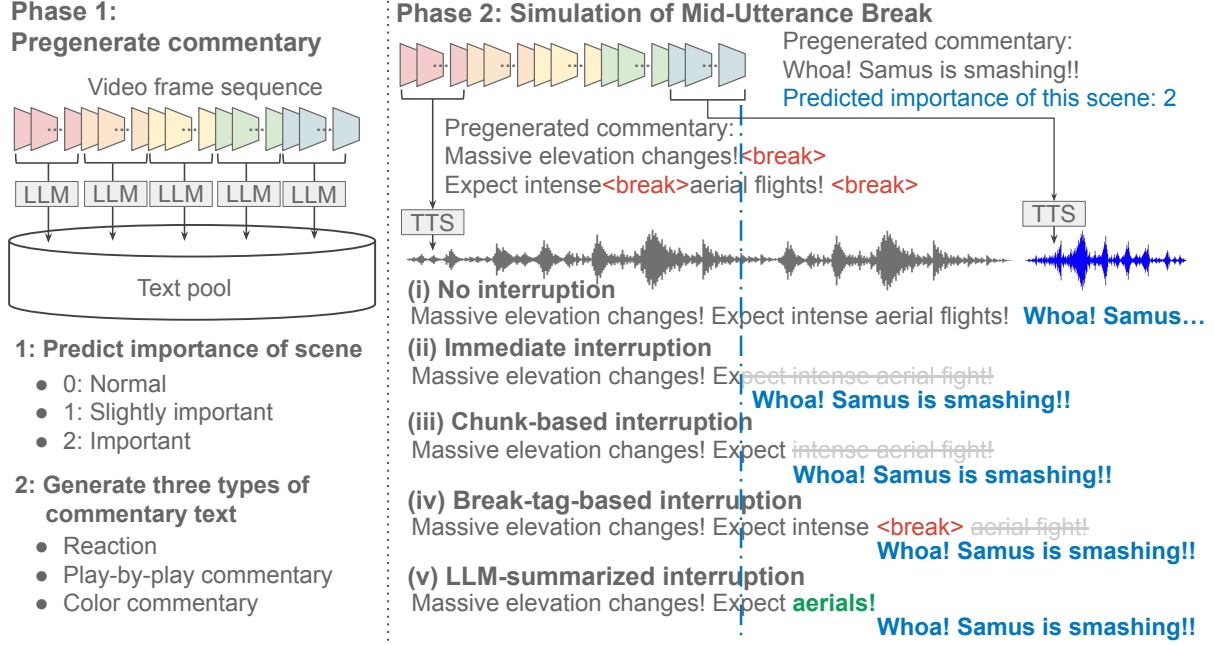


Figure 2: Overview of the two-phase evaluation framework for simulating mid-utterance transitions. To ensure a fair comparison, Phase 1 processes video frame sequences via LLMs to predict scene importance scores (0–2) and pre-generates diverse commentary candidates to construct a fixed text pool. In Phase 2, the system simulates live playback by synthesizing speech from this pool in sync with the gameplay timeline. When a high-importance event (score 2) occurs during an ongoing utterance, the simulation handles the transition by either completing the current speech (i) or interrupting it mid-utterance ((ii)–(v)). All five strategies are evaluated both with and without an interjection.

available and immediately begins playing the new utterance. This strategy minimizes response latency, but the stopping point is determined solely by the event timing. Consequently, the audio may be truncated in the middle of a word or syntactic unit, resulting in both acoustic discontinuity and grammatical incompleteness. This strategy has been used in an existing commentary generation system (Tanaka-Ishii et al., 1998; Ishigaki et al., 2023).

Chunk-based interruption: The system delays the transition until the nearest upcoming syntactic chunk boundary in C_k . We use Japanese syntactic chunks boundaries (i.e., *bunsetsu*) obtained through a morphological analyzer (McCann, 2020). When an interruption is triggered, playback continues from its current position to the end of the current chunk and then transitions to C_{k+m} . This prevents mid-word truncation and introduces only a short additional delay.

Break-tag-based interruption: This strategy identifies potentially natural interruption points during commentary generation. The LLM is prompted to insert a <break> tag at positions where the utterance could be suspended without severe semantic

or grammatical disruption. When an interruption is triggered, the system continues playback until the next tagged position and then starts C_{k+m} . Unlike the chunk-based strategy, which selects boundaries using local part-of-speech patterns, this strategy allows the LLM to consider the entire utterance when choosing candidate stopping points. The retained prefix may therefore sound less semantically or grammatically incomplete, although waiting for the next tag can increase response latency.

LLM-summarized interruption: This strategy revises the remaining portion of the ongoing utterance rather than waiting for its originally planned ending. When an interruption is triggered, the system first retains the portion of C_k that has already been spoken up to a syntactic boundary. An LLM then generates a short continuation that grammatically closes or redirects this spoken prefix. The generated closure is synthesized and played before the system transitions to C_{k+m} . By replacing the unspoken remainder of C_k with a shorter completion, this strategy aims to preserve linguistic coherence while reacting more quickly than a system that plays the entire original utterance.

Following the human-commentary observations,

Method	Response Latency [s]
Baseline	2.963 (± 1.198)
Immediate	0.080 (± 0.027)
Chunk-based	0.327 (± 0.210)
Break-tag	1.381 (± 0.897)
Summarized	1.774 (± 0.514)

Table 2: Response latency (time from event trigger to reaction onset) for each method, including text-to-speech synthesis latency (0.080 [s]). Values in parentheses indicate standard deviation.

we additionally evaluate each transition strategy both with and without an interjection inserted immediately before C_{k+m} . The interjection is intended to signal the sudden change in the game situation and to reduce the perceived abruptness of the transition.

5 Experiments

We describe the experimental setup to compare five transition strategies. Different transition methods stop the ongoing speech at slightly different points. In the online setting described in Section 3, these differences would accumulate, potentially causing each method to eventually interrupt completely different commentary utterances. To ensure a strictly fair comparison on the exact same utterances, we evaluated the strategies on a fixed set of pre-generated utterance pairs, as described in Figure 2.

5.1 Dataset

We used the SMASH corpus (Saito et al., 2020), which contains gameplay videos and human commentary audio from the competitive fighting game *Super Smash Bros. Ultimate*. The corpus includes both Singles matches, involving two players, and Doubles matches, involving two teams of two players. To cover different match configurations, we randomly selected four Singles matches and four Doubles matches, yielding eight gameplay videos in total.

5.2 Compared Methods

To examine the trade-off between responsiveness and acoustic and linguistic naturalness, we compared four interruption strategies with a non-interruptive baseline, which plays the ongoing utterance to completion before beginning the newly selected utterance.

5.3 Construction of Evaluation Instances

For each selected video, we generated the three commentary types described in Section 3.1 at one-second intervals. As depicted in Phase 1 of Figure 2, we prompted the LLM to concurrently predict an importance score of 0, 1, or 2 for each interval. We constructed a candidate interruption instance by pairing a score-2 commentary with an earlier utterance whose synthesized audio was still playing at the score-2 timestamp. Each instance therefore represents a situation in which a newly detected important event must be reported before the ongoing utterance has finished.

Because our goal is to evaluate transition quality rather than commentary generation quality, we manually filtered these candidates. We sampled 40 instances in which both utterances were factually consistent with the gameplay and sufficiently natural, preventing generation errors from influencing judgments of the transition itself. For these instances, we applied all five transition strategies, both with and without an interjection, producing 400 evaluation samples in total (40 instances $\times$ 5 transition methods $\times$ 2 patterns of interjection).

5.4 Evaluation Metrics

We evaluated the transition strategies using both objective and subjective measures.

Response Latency: We measured the elapsed time from the detection of an event with the highest importance score of 2 to the onset of audio playback for the newly selected commentary. This metric quantifies how quickly each strategy begins reporting an important game event.

Manual Evaluation by Human Judges: Each of the 400 test instances was evaluated by two independent Japanese native speakers (not the authors) who have experience playing *Super Smash Bros*. For full details regarding the evaluator profiles and the exact instructions provided, please refer to Appendix B. They rated each sample on a five-point Likert scale for the following criteria:

- **Responsiveness:** How promptly did the commentary react to the important game event?
- **Linguistic Naturalness:** How natural was the linguistic transition from the ongoing utterance to the new one?
- **Acoustic Naturalness:** How natural was the audio transition between the two utterances?

Method	Responsiveness	Linguistic Nat.	Acoustic Nat.	Overall Quality
Baseline	1.86 (± 0.27)	4.09 (± 0.20)	3.66 (± 0.24)	2.66 (± 0.20)
+ Interjection	1.77 (± 0.27)	3.75 (± 0.25)	3.59 (± 0.22)	2.66 (± 0.18)
Immediate	4.01 (± 0.25)	2.38 (± 0.30)	2.19 (± 0.26)	2.75 (± 0.22)
+ Interjection	4.12 (± 0.25)	2.66 (± 0.29)	2.75 (± 0.31)	3.10 (± 0.25)
Chunk-based	3.59 (± 0.27)	2.85 (± 0.26)	2.88 (± 0.26)	3.00 (± 0.21)
+ Interjection	3.44 (± 0.31)	3.23 (± 0.27)	3.17 (± 0.27)	3.19 (± 0.21)
Break-tag	2.48 (± 0.32)	3.79 (± 0.21)	3.41 (± 0.24)	2.94 (± 0.20)
+ Interjection	2.34 (± 0.33)	3.59 (± 0.28)	3.33 (± 0.27)	2.91 (± 0.22)
Summarized	2.11 (± 0.25)	3.30 (± 0.25)	2.95 (± 0.26)	2.54 (± 0.17)
+ Interjection	2.04 (± 0.26)	3.33 (± 0.27)	3.04 (± 0.25)	2.61 (± 0.18)

Table 3: Subjective evaluation results. Values in parentheses indicate $\pm$ half-width of the 95% confidence interval. The results demonstrate a clear trade-off: immediate interruption maximizes responsiveness but severely degrades naturalness, whereas chunk-based interruption with an interjection provides the most effective overall balance.

- **Overall Quality:** What was the overall quality of the resulting commentary?

We report the mean opinion score (MOS) for each criterion.

5.5 Implementation Details

For each one-second interval, 30 uniformly sampled frames were encoded in Base64 and provided to GPT-4.1². The prompts used for commentary generation are provided in Appendix A.1.

We used the commentary-specific Japanese text-to-speech Iura et al. (2025). The model allows the speaking style of each utterance to be controlled using a binary excitement parameter, *High* or *Low*. The *High* setting produces an excited delivery suitable for decisive events, whereas *Low* produces a calmer delivery. We therefore set the excitement level to *High* for utterances associated with score-2 events and to *Low* for those associated with scores 0 or 1. Text-to-speech inference was performed on a single NVIDIA A30 GPU.

6 Results

6.1 Response Latency

As shown in Table 2, mid-utterance transitions reduce latency compared with the baseline. The immediate method achieved the shortest latency at 0.080 seconds, followed by the chunk-based method at 0.327 seconds. The LLM-based methods incurred longer delays of 1.381 and 1.774 seconds, respectively for the break tag-based and LLM-Summarized methods, although both remained faster than the baseline at 2.963 seconds. Note

that these values include text-to-speech synthesis latency, which averaged 0.080 seconds across all utterances. Overall, mid-utterance strategies has advantage of responsiveness. We discuss the tradeoff between responsiveness and naturalness next.

6.2 Subjective Evaluation

Table 3 suggests a clear trade-off between responsiveness and naturalness: faster strategies tended to produce less natural transitions, while more natural strategies tended to respond more slowly.

In terms of responsiveness, scores showed similar tendencies to the scores of automatic evaluation. The immediate interruption achieved the highest responsiveness score of 4.01, followed by the chunk-based interruption at 3.59, the break-tag-based interruption at 2.48, LLM-summarized interruption at 2.11, and the Baseline at 1.86.

The scores in terms of naturalness showed the opposite tendencies, which suggest the tradeoff. The Baseline received the highest linguistic-naturalness score of 4.09, followed by the break-tag-based interruption at 3.79. The immediate interruption received the lowest linguistic- and acoustic-naturalness scores, at 2.38 and 2.19, respectively. These results show that completing the ongoing utterance preserves naturalness, whereas immediately truncating it maximizes responsiveness at a substantial cost to linguistic and acoustic quality.

In terms of the overall quality, evaluators gave the highest score for the chunk-based interruption (3.00), followed by the break-tag-based interruption (2.94). Possible reason behind this may be the better balance between linguistic completeness and responsiveness.

²<https://developers.openai.com/api/docs/models/gpt-4.1>

Baseline

R.O.B. is at a very high 148% damage, but a launch here is a critical moment that will heavily impact the stocks. The recovery after this is also an important mind game.
 $\Rightarrow$ “Both are launched away at once!”

Immediate

R.O.B. is at a v†
 $\Rightarrow$ “Both are launched away at once!”

Chunk-Based

R.O.B. is at a very
 $\Rightarrow$ “Both are launched away at once!”

Break-Tag

R.O.B. is at a very high 148% damage, but a launch here
 $\Rightarrow$ “Both are launched away at once!”

Summarized

*R.O.B. is at a very high 148% damage, but **that’s dangerous.***
 $\Rightarrow$ “Both are launched away at once!”

Table 4: Example outputs of each interruption method. Each block displays the pre-interrupt utterance followed by the subsequent reaction ($\Rightarrow$). † indicates a mid-syllable interruption, and bold text represents a dynamically generated closing phrase. Texts are translated from Japanese.

6.3 Effect of Interjections

The effect of interjections depended on how complete the interrupted utterance sounded at the transition point. For the immediate and chunk-based interruption, the ongoing utterance may be cut at a semantically incomplete position. In these cases, an interjection explicitly signals a sudden shift in attention and helps listeners interpret the abrupt change as an intentional transition rather than an error. Consistent with this interpretation, adding an interjection significantly improved acoustic naturalness from 2.19 to 2.75 ($p < 0.001$) and overall quality from 2.75 to 3.10 ($p < 0.01$) for the immediate interruption (Wilcoxon signed-rank test). For the chunk-based interruption, linguistic naturalness increased from 2.85 to 3.23 and acoustic naturalness from 2.88 to 3.17, both of which were statistically significant (Wilcoxon signed-rank test; $p < 0.01$ and $p < 0.05$, respectively). The interjection condition achieved the highest overall-quality score in the study, 3.19, although this improvement did not reach statistical significance.

By contrast, the LLM-based strategies are designed to stop at a more semantically self-contained position or to complete the interrupted utterance before switching topics. As a result, the transition can already sound coherent without an explicit discourse signal. For the break-tag-based interrup-

tion, adding an interjection reduced all four subjective scores, suggesting that the additional cue was perceived as redundant or disruptive. For LLM-summarized interruption, the effect was small, with overall quality increasing only from 2.54 to 2.61.

Overall, interjections appear most useful when the retained prefix is semantically incomplete. They can mark the topic shift and mitigate the abruptness of the interruption. When the stopping point is already semantically natural, however, an interjection provides little additional benefit and may even reduce naturalness.

6.4 Qualitative Analysis

6.4.1 Examples of Transition Behaviors

Table 4 illustrates actual outputs, revealing that while avoiding mid-word truncation is necessary, fully completing the interrupted utterance is not always beneficial.

The immediate interruption stops mid-word (“v†”), sounding like a playback error rather than an intentional reaction. The chunk-based interruption avoids this by waiting for the next boundary. Although the retained prefix (“R.O.B. is at a very”) remains semantically incomplete, the boundary provides a natural pivot to the new event, making the transition sound intentional.

The break-tag-based interruption retains a more self-contained phrase, and LLM-summarized interruption generates a grammatical closure (“that’s dangerous.”). However, both methods inherently delay the reaction to the new event, which can sound unnecessary in situations requiring an immediate response.

6.5 Evaluators’ Feedback

Post-evaluation interviews with evaluators revealed two key insights.

First, evaluators were relatively tolerant of minor linguistic and acoustic disruptions when the interruption was timely. Highly responsive mid-sentence interruptions enhanced immersion, as they were perceived as intentional, human-like reactions rather than generation errors.

Second, delayed emotional reactions were perceived as particularly unnatural; for instance, a surprised tone sounded jarring if the triggering event had already concluded. Consistent with our quantitative results, interjections improved transitions for faster strategies but provided little benefit for slower ones, indicating they are most effective when temporally aligned with the event.

7 Conclusion

We investigated mid-utterance transition strategies for live audio commentary generation in fast-paced games. Rather than treating speech playback as an atomic process, we formulated the problem as deciding when to terminate an ongoing utterance and how to transition to a newly selected higher-priority utterance. We compared four interruption strategies—the immediate, chunk-based, break-tag-based, and LLM-summarized interruption—against a non-interruptive baseline, with and without interjections. Our results reveal a clear trade-off between responsiveness and naturalness among compared methods. The qualitative feedback further suggests that listeners tolerate minor disfluencies when interruptions are timely, while delayed emotional reactions can sound particularly unnatural. Our results suggest mid-utterance transition control as an important component in future developments of automated commentary systems.

Limitations

While this work demonstrates the importance of mid-utterance transition control, several limitations remain. First, our evaluation was conducted in a simulated environment rather than a fully deployed real-time system. Second, this research focuses exclusively on Super Smash Bros., a competitive fighting game; further research is needed to determine whether the findings can be generalized to other game genres or languages.

Ethical Considerations

This research uses LLMs, which carry inherent risks of occasionally generating hallucinations or offensive language despite prompt constraints. Regarding the human evaluation, we did not collect any personally identifiable information. The task of rating game commentaries poses minimal medical or psychological risks, and all evaluators received a clear explanation of the task beforehand.

Acknowledgments

This paper is based on results obtained from AIST policy-based budget project "R&D on Generative AI Foundation Models for the Physical Domain. This work was primarily conducted while the second author was a researcher at AIST. His current primary affiliation is Sports AI Tech Lab, CyberAgent, Inc.

References

- Anum Afzal, Yuki Saito, Hiroya Takamura, Katsuhito Sudoh, Shinnosuke Takamichi, Graham Neubig, Florian Matthes, and Tatsuya Ishigaki. 2026. Real-time generation of game video commentary with multi-modal llms: Pause-aware decoding approaches. In *Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC)*, pages 9188–9201.
- Anton Behrens and Sebastian Uhrich. 2022. 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.
- Okko Buß and David Schlangen. 2011. DIUM – an incremental dialogue manager. In *Proceedings of the 15th Workshop on the Semantics and Pragmatics of Dialogue (SemDial)*.
- Joya Chen, Ziyun Zeng, Yiqi Lin, Wei Li, Zejun Ma, and Mike Zheng Shou. 2025. LiveCC: Learning video llm with streaming speech transcription at scale. In *Proceedings of the 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 29083–29095.
- Jaime A. Teixeira da Silva and Nicolas Scelles. 2025. A vision for the formal documentation and digitalization of sports commentators’ commentaries. *International Journal of Sport Communication*, 18(1):1–9.
- Alexandre Défossez, Laurent Mazaré, Manu Orsini, Amélie Royer, Patrick Pérez, Hervé Jégou, Edouard Grave, and Neil Zeghidour. 2024. [Moshi: a speech-text foundation model for real-time dialogue](#). Preprint, arXiv:2410.00037.
- Rebecca Heins, Marita Franzke, Michael Durian, and Aruna Bayya. 1997. Turn-taking as a design principle for barge-in in spoken language systems. *International Journal of Speech Technology*, 2(2):155–164.
- Julian Hough. 2011. Incremental semantics driven natural language generation with self-repairing capability. In *Proceedings of the Second Student Research Workshop associated with Recent Advances in Natural Language Processing (RANLP)*, pages 79–84.
- Tatsuya Ishigaki, Goran Topic, Yumi Hamazono, Ichiro Kobayashi, Yusuke Miyao, and Hiroya Takamura. 2023. Audio commentary system for real-time racing game play. In *Proceedings of the 16th International Natural Language Generation Conference (INLG)*, pages 9–10.
- Tatsuya Ishigaki, Goran Topic, Yumi Hamazono, Hiroshi Noji, Ichiro Kobayashi, Yusuke Miyao, and Hiroya Takamura. 2021. Generating racing game commentary from vision, language, and structured data. In *Proceedings of the 14th International Conference on Natural Language Generation (INLG)*, pages 103–113.

- Kota Iura, Yuki Saito, Shinnosuke Takamichi, Graham Neubig, Katsuhito Sudoh, Hiroshi Saruwatari, Hiroya Takamura, and Tatsuya Ishigaki. 2025. Excitement-inducing commentary text-to-speech system for fighting game video scenes. *IEEE Access*, 13:216748–216758.
- Ryota Kawamatsu, Anum Afzal, Yuki Saito, Shinnosuke Takamichi, Graham Neubig, Katsuhito Sudoh, Hiroya Takamura, and Tatsuya Ishigaki. 2026. [Low-latency real-time audio game commentary system via llm-based parallel text generation](#). *Preprint*, arXiv:2606.13322.
- Byeong Jo Kim and Yong Suk Choi. 2020. Automatic baseball commentary generation using deep learning. In *Proceedings of the 35th Annual ACM Symposium on Applied Computing (SAC)*, pages 1056–1065.
- Paul McCann. 2020. `fugashi`, a tool for tokenizing Japanese in python. In *Proceedings of the Second Workshop for NLP Open Source Software (NLP-OSS)*, pages 44–51.
- Yuichiro Mori, Chikara Tanaka, Aru Maekawa, Satoshi Kosugi, Tatsuya Ishigaki, Kotaro Funakoshi, Hiroya Takamura, and Manabu Okumura. 2025. Live football commentary system providing background information. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 394–404.
- Jiayuan Rao, Haoning Wu, Chang Liu, Yanfeng Wang, and Weidi Xie. 2024. MatchTime: Towards automatic soccer game commentary generation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1671–1685.
- Yuki Saito, Shinnosuke Takamichi, and Hiroshi Saruwatari. 2020. SMASH corpus: A spontaneous speech corpus recording third-person audio commentaries on gameplay. In *Proceedings of the Twelfth Language Resources and Evaluation Conference (LREC)*, pages 6571–6577.
- Gabriel Skantze. 2021. Turn-taking in conversational systems and human-robot interaction: A review. *Computer Speech & Language*, 67:101178.
- Taiga Someya, Tatsuya Ishigaki, and Hiroya Takamura. 2025. Live football commentary (LFC): A Large-Scale dataset for building football commentary generation models. In *Proceedings of the 18th International Natural Language Generation Conference (INLG)*, pages 195–201.
- Nikko Ström and Stephanie Seneff. 2000. Intelligent barge-in in conversational systems. In *Proceedings of the 6th International Conference on Spoken Language Processing (ICSLP)*, pages vol. 2, 652–655.
- Haoran Sun, Natthawut Kertkeidkachorn, and Kiyooki Shirai. 2026. Visual and memory-augmented soccer commentary generation. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 10614–10629.
- Kumiko Tanaka-Ishii, Koiti Hasida, and Itsuki Noda. 1998. Reactive content selection in the generation of real-time soccer commentary. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics (COLING-ACL)*, page 1282–1288.
- Yasufumi Taniguchi, Yukun Feng, Hiroya Takamura, and Manabu Okumura. 2019. Generating live soccer-match commentary from play data. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, pages 7096–7103.
- Bandhav Veluri, Benjamin N Peloquin, Bokai Yu, Hongyu Gong, and Shyamnath Gollakota. 2024. Beyond turn-based interfaces: Synchronous LLMs as full-duplex dialogue agents. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 21390–21402.
- Zihan Wang and Naoki Yoshinaga. 2025. 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 (IJCNLP-AAACL)*, pages 1795–1807.
- Mirjam Wester, David A. Braude, Blaise Potard, Matthew P. Aylett, and Francesca Shaw. 2017. Real-Time Reactive Speech Synthesis: Incorporating Interruptions. In *Proceedings of the Interspeech 2017 (Interspeech)*, pages 3996–4000.
- Huanyu Yu, Shuo Cheng, Bingbing Ni, Minsi Wang, Jian Zhang, and Xiaokang Yang. 2018. Fine-grained video captioning for sports narrative. In *Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6006–6015.
- Qirui Zheng, Xingbo Wang, Keyuan Cheng, Muhammad Asif Ali, Yunlong Lu, and Wenxin Li. 2025. [From multimodal perception to strategic reasoning: A survey on ai-generated game commentary](#). *Preprint*, arXiv:2506.17294.
- Xingyi Zhou, Anurag Arnab, Shyamal Buch, Shen Yan, Austin Myers, Xuehan Xiong, Arsha Nagrani, and Cordelia Schmid. 2024. Streaming dense video captioning. In *Proceedings of the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18243–18252.

A Prompts Used in Experiments

A.1 Commentary Generation Prompt

The following prompt was used to generate `play_by_play` and `color_commentary` texts for each video window. The model receives a sequence

Metrics	Fleiss's κ	Cohen's κ	Krippendorff's α
Responsiveness	0.513	0.636	0.619
Linguistic Nat.	0.160	0.202	0.194
Acoustic Nat.	0.095	0.134	0.099
Overall Quality	0.158	0.232	0.200

Table 5: Inter-annotator agreement between two fully-rated evaluators. Fleiss's κ : Cohen's Kappa; α : Krippendorff's Alpha.

of video frames alongside match context, and outputs a JSON object including the two commentary fields with <break> markers indicating safe interruption points.

あなたは大乱闘スマッシュブラザーズのプロ実況者です。

```
## 試合情報
- 形式: {match_format_label}
{teams_block}
## 前窓のHUD状態
{hud_block}
{ko_block}
## これまでの実況（直近
{context_windows} 窓）
{past_pbp_block}
-
以下は試合映像の {n_frames} フレーム
（約 {window_sec} 秒）です。
{phase_block}

## ハルシネーション防止ルール（必ず守る
こと）
- 映像から視覚的に確認できることだけを
描写する
- 「命中した」「ダメージを与えた」「当
たった」など結果・判定を断定しない
→ 「～のモーションを取っている」「～に
向かって技を出した」のように表現する
- キャラクターの識別が不確かな場合は名
前を使わず「左のキャラ」等にする
- ダブルスでチーム帰属が不明なキャラへ
の言及は避ける
- 吹き飛んでいるキャラを「撃墜」とは言
わない（ステージ外に消えるまで確認で
きない）
- HUD に表示されていない数値は書かない
```

```
## importance の仕様
この値は play_by_play と
color_commentary のどちらをどのタ
イミングで読み上げるかを決定します:
- 0: 通常シーン。play_by_play と
color_commentary が均等に選ばれます。
- 1: やや重要なシーン。play_by_play が
優先して選ばれます（80%）。
- 2: 重要【割り込みトリガー】進行中の発
話を中断し、play_by_play を即座に割り込
ませます。
中断してでも視聴者に伝えるべき瞬間にの
み選んでください。
```

（出力は JSON 形式。importance
/ importance_reason / reaction /
play_by_play / color_commentary /

game_phase / hud_state / inferred_teams
を含む。）

<break> は「音声我突然止まっても不自然
でない位置」に挿入してください。
読点（、）や句点（。）の直後は避けてく
ださい。

A.2 Ender Generation Prompt

The following prompt was used in the **Summa-
rized** method to generate a short closing phrase
that naturally completes the interrupted utterance
at a bunsetsu boundary. The system prompt is

実況者が以下のテキストを喋っている途中
で発話を中断します：
{full_text}

現在「{stub}」まで喋ったところで中断し
ます。

「{stub}」の直後で文を自然に短く締める
テキストを1つだけ生成してください。

条件：

- 「{stub}」に文法的に接続する
（「{stub}」自体は出力しない）
- できるだけ短く（0～8文字）
- 感嘆符（!）または句点（。）で終わる
- 新しい情報を追加せず、文を自然に閉じ
るだけ

締めめのテキストのみを出力してください。
説明・引用符不要。

B Details of Human Evaluation

Evaluator Profiles: To ensure a fair and high-
quality assessment, the human evaluation was
conducted by two independent annotators who
were not involved in the authorship of this paper.
Both evaluators are native Japanese speakers and
have prior experience playing and watching Super
Smash Bros., ensuring they possess the necessary
domain knowledge to properly judge the context,
timing, and naturalness of the game commentary.

Evaluation Instructions and Interface: Since
the generated commentary and the source game
videos were in Japanese, the evaluation interface
and all instructions were provided to the evaluators
in their native language (Japanese). Figure 3 shows

a screenshot of the actual evaluation interface. The evaluators were instructed to watch each video clip with the generated audio and rate the commentary on a 1-to-5 Likert scale across four dimensions.

The English translations of the exact instructions provided to the evaluators are as follows:

- **Responsiveness:** How promptly did the commentary react to the important game event?
- **Linguistic Naturalness:** How natural was the linguistic transition from the ongoing utterance to the new one?
- **Acoustic Naturalness:** How natural was the audio transition between the two utterances?
- **Overall Quality:** What was the overall quality of the resulting commentary?

C Inter-Annotator Agreement

Table 5 reports inter-annotator agreement between the two fully-rated evaluators. Krippendorff’s α (ordinal) was 0.619 for responsiveness, indicating moderate agreement, whereas linguistic naturalness, acoustic naturalness, and overall quality showed lower agreement ($\alpha = 0.10$ – 0.20). This pattern is consistent with prior work on subjective speech evaluation, reflecting the inherent subjectivity of naturalness perception.

Specifically, absolute agreement metrics like Krippendorff’s α heavily penalize differences in evaluators’ internal baselines. For instance, a transition that Evaluator A rates as a “3” might be consistently rated as a “2” by Evaluator B due to differing levels of strictness. While this systematic shift results in low absolute agreement scores, it does not undermine our conclusions: as long as both evaluators agree on the relative ranking of the strategies (e.g., both agree that strategy Y is more natural than strategy X), the consensus on which method performs better remains intact.

AI ゲーム実況 評価タスク

セット 14 / 20

(動画 128 / 400)

0:01

0:06

1

非常に低い

2

低い

3

普通

4

高い

5

非常に高い

全く感じられなかった

強く感じられた

1

非常に低い

2

低い

3

普通

4

高い

5

非常に高い

非常に不自然

非常に自然

1

非常に低い

2

低い

3

普通

4

高い

5

非常に高い

非常に不自然

非常に自然

1

非常に低い

2

低い

3

普通

4

高い

5

非常に高い

非常に低い

非常に高い

1

非常に低い

2

低い

3

普通

4

高い

5

非常に高い

非常に低い

非常に高い

途中保存してダウンロード

次へ →

Figure 3: Screenshot of the human evaluation interface. The interface was presented in Japanese, matching the language of the generated commentary and the native language of the evaluators. Evaluators rated each sample on a 5-point Likert scale across four dimensions (from top to bottom): Responsiveness, Linguistic Naturalness, Acoustic Naturalness, and Overall Quality.

695