Measuring Time Delay Tolerance in Third-Person Live Commentary for Super Smash Bros. Ultimate

Ryosuke Matsushita
Keio University, Japan
ryosuke.jp66@keio.jp

Ryosuke Sakai
Keio University, Japan

Koki Fukuda
Keio University, Japan

Shinnosuke Takamichi
Keio University, Japan
shinnosuke_takamichi@keio.jp

Kota Iura
The University of Tokyo, Japan

Yuki Saito
The University of Tokyo, Japan

Graham Neubig
Carnegie Mellon University, U.S.A.

Katsuhito Sudoh
Nara Women's University, Japan

Hiroya Takamura
National Institute of Advanced
Industrial Science and Technology, Japan

Tatsuya Ishigaki
National Institute of Advanced
Industrial Science and Technology, Japan

***Abstract*—This study proposes a methodology for measuring the acceptable delay tolerance for third-person game commentary. Third-person game commentary refers to commentary delivered by someone other than the player, with the role of helping viewers better understand the game and enhancing the viewing experience. With the recent advancement of AI, there has been increasing interest in automating such commentary using video understanding and audio generation. However, automating this process using video understanding and audio generation introduces delays, potentially affecting the naturalness of the commentary. In this context, since the extent to which such delays are acceptable to viewers remains unclear, we address this issue. The tolerance is modeled using an unnormalized Gaussian function. Through experiments on Super Smash Bros. Ultimate with 727 participants, we found that the average acceptable delay for this game is 3.71 seconds, with variations depending on different viewer attributes and gameplay contexts.**

***Index Terms*—game commentary, delay tolerance, gameplay**

I. Introduction

A third-person game commentary refers to a style of game commentary where a commentator explains the game content while conveying the progress of the game to the players or game viewers. The commentator observes the game screen and actions taken by the player(s) and verbally describes them, serving to convey the gameplay scenes in an easily understandable manner to the viewers [1]. Timely delivery is crucial for effective commentary. In both television broadcasts and online video streaming, even slight delays between the in-game action and the corresponding commentary can reduce the naturalness and quality of the viewing experience. Despite its importance, providing high-quality commentary requires specialized skills, and as a result, many gameplay videos uploaded online do not include commentary [2].

Given this context, researchers have developed methods to automatically generate commentary from gameplay videos [2]–[6]. However, when automatically generating commentary, the processes of video understanding [7], [8] and audio generation [9], [10] take time, resulting in delays in the generated audio commentary. To provide natural audio commentary to viewers, the process must be completed within a time frame acceptable to the viewer. Since the link between commentary delay and perceived naturalness has not been explored, this study provides new insight.

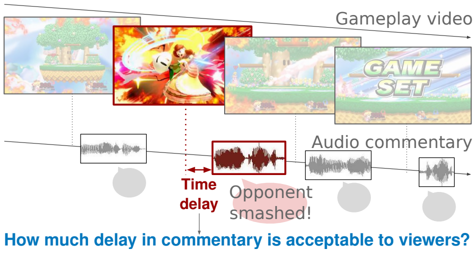

Fig. 1. Example of delay of audio commentary for Super Smash Bros. Ultimate.

Therefore, this study proposes a methodology to measure the acceptable delay in third-person game commentary. Specifically, as shown in Figure 1, we conduct experiments using gameplay videos to evaluate the extent to which viewers feel discomfort due to timing mismatches.

Using Super Smash Bros. Ultimate[1] as the target game, we investigate the acceptable delay and the effects of the viewer attributes and gameplay content. The results revealed that an average delay of 3.71 seconds is acceptable, with variation across viewer and gameplay factors.

II. Related work

A. Subjective evaluation of delays

Previous studies on delay measurement [11]–[13] have employed the five-point method of constant stimuli [14]. The method is a technique where participants make subjective judgments about given items based on a predefined five-level evaluation. This method is also adopted in this study.

[1] https://www.smashbros.com/en_US/index.html

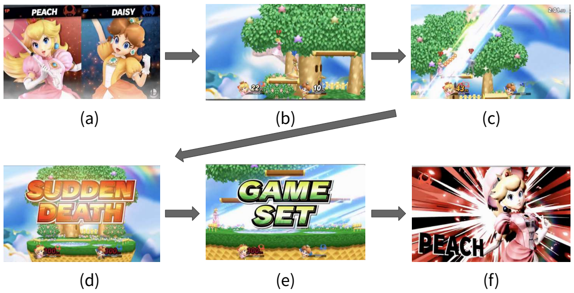

Fig. 2. Flow of gameplay video.

B. Effects of viewer attributes

It has been shown that response times to text or audio stimuli vary depending on participants' attributes [15], [16]. Although there are no studies on the response time (acceptable delay) for listening to game commentary audio, it is expected that viewer attributes have an influence. Therefore, this study also collects data on viewer attributes.

III. Proposed Method

We propose a method to measure the acceptable delay for third-person game commentary audio. It uses a dataset consisting of time-aligned gameplay videos and third-person game commentary audio. The commentary audio is annotated with the start time and a topic tag for each utterance. The topic tag represents a label for the commentary content of the utterance, e.g., battle scenes or results. Based on this dataset, video stimulus sets to be presented to participants are first created. Then, the video stimuli are presented to many participants, who are asked to evaluate whether the delay is perceived as unnatural or natural. Their responses are aggregated to estimate acceptable delay. Delay values include negative cases, where commentary starts before the reference time, enabling evaluation of tolerance for early commentary.

A. Super Smash Bros. Ultimate

Super Smash Bros. Ultimate is a fast-paced game where delay issues can be particularly severe. As shown in Figure 2, the game consists of several phases. In phase (a), the characters are introduced. In phases (b) and (c), battles take place. If the scores are tied within the time limit, a sudden-death phase (d) is conducted. Following this, the match ends (e), and the results are announced (f). The dataset used is an existing corpus that includes commentary audio [17].

B. Creation of video stimulus

To evaluate the acceptable delay of the commentary audio, we created video segments from the dataset. Each video segment consists of the commentary audio, including the target utterance and the preceding N utterances, as well as the corresponding gameplay video. Including the preceding N utterances in addition to the target utterance is intended to help participants understand the context and status of the gameplay. The value of N is set to the maximum possible value such that the duration of the video segment does not exceed a predefined threshold (the maximum duration). The analysis of the evaluation results is based on the topic tag assigned to the final utterance. A time delay is then applied to the commentary audio. The candidates for delay amounts are predefined, and video stimuli with all combinations of video segments and delay amount candidates are created. The delay amount candidates include positive, negative, and 0. The 0 serves as a reference, i.e., no delay is applied.

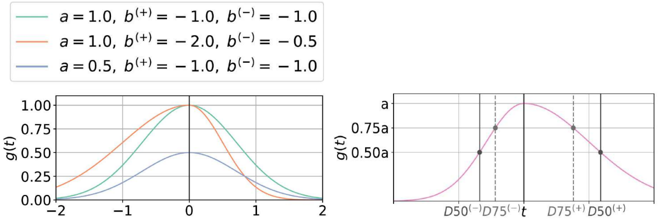

Fig. 3. Regression curves (left) and acceptable delay times (right).

C. Creating balanced stimulus sets for each participant

We created video stimulus sets to be presented to each participant. When creating these sets, it is necessary to ensure uniformity to prevent unintended biases from affecting the evaluations due to the combination of participants and sets. This study ensures uniformity in the following two aspects:

- **Delay amounts**: every stimulus set is constructed to include an equal number of videos for each candidate delay value to ensure participants evaluated a balanced range of delay conditions. This guarantees that no participant evaluates only a biased subset, such as only large positive delays.
- **Utterance topics**: It is expected that commentary is required for sudden in-game events, while greater delays may be acceptable for chat. Therefore, it is necessary that the stimulus sets are not biased toward specific topics. To achieve this, the frequency distribution of utterance topics in each video stimulus set is calculated and compared to the overall distribution in the dataset. All video stimulus sets are created so that the Kullback–Leibler divergence between the two distributions does not exceed a predefined threshold. A greedy algorithm [18] is used for the process.

As described later, the evaluation is based on the method of constant stimuli using a five-point scale [14], so the order of video stimuli within each set is randomized.

D. Evaluation of time delay by participants

Each participant views the video stimuli and performs a five-level MOS (mean opinion score) evaluation of the temporal misalignment between the gameplay video and the commentary audio, using a scale from 1 (very unnatural) – 5 (very natural). After evaluating all video stimuli, participants complete a questionnaire about their attributes.

E. Measurement of acceptable delay time

The evaluation scores are averaged for each delay time, and the following regression curve is fitted to the data:

$$g(t) = \begin{cases} a \cdot \exp(b^{(+)} \cdot t^2) & (t \geq 0) \\ a \cdot \exp(b^{(-)} \cdot t^2) & (t < 0) \end{cases} . \tag{1}$$

Here, t represents the delay time. The general shape of this equation is illustrated in Figure 3. This function is a non-normalized Gaussian function with different precision parameters on either side of $t = 0$. $a \geq 0$ is a scale parameter, and $b^{(-)} < 0,\ b^{(+)} < 0$ are the negative precision parameters.

In this paper, the acceptable delay times are defined as the durations where the value of the regression curve reaches 75% and 50% of its maximum value (Figure 3). The acceptable delay times, denoted as $\mathrm{D50}^{(-)}, \mathrm{D75}^{(-)}, \mathrm{D75}^{(+)}, \mathrm{D50}^{(+)}$ (where $^{(+)}$ and $^{(-)}$ correspond to positive and negative delays, respectively), are calculated as follows:

$$\mathrm{D50}^{(\pm)} = \sqrt{\frac{\ln 0.5}{b^{(\pm)}}},\ \mathrm{D75}^{(\pm)} = \sqrt{\frac{\ln 0.75}{b^{(\pm)}}}. \qquad (2)$$

IV. Experimental evaluation

A. Experimental conditions

We measured the acceptable delay for commentary audio in Super Smash Bros. Ultimate. Participants were recruited through the crowdsourcing platform Lancers[2], totaling 727 individuals. Each participant was compensated 240 JPY.

1) Creation of video stimuli and stimulus sets: The video stimuli were created using 32 one-on-one match videos, totaling approximately 1.3 hours, from the SMASH Corpus [17]. This corpus comprises gameplay videos, commentary audio, topic tags, and utterance start times from the game Super Smash Bros. Ultimate. The matches followed a time-limited battle format of 150 seconds and included the following sections: character introductions at the beginning, a 150 - second match in the middle, and match results at the end. The following topic tags were evaluated in this study: commentaries on 1) match results (Result), 2) match scenes (Scene), 3) items appearing during the match (Item), and 4) the characters in the match (Fighter). Utterances with chat topics, which are not directly related to the match, were excluded as they are irrelevant for delay evaluation. Commentary was provided by two Japanese male commentators (one per video). Audio was sampled at 16 kHz; video ran at 30 fps.

The maximum duration of the video stimuli was set to 20 seconds, a value determined through a preliminary experiment to allow viewers to grasp the match situation. The candidate delay amounts ranged from -1.50 seconds to 1.75 seconds, in increments of 0.25 seconds, for a total of 12 options. A total of 5,016 video stimuli were finally created. Each participant was presented with 24 video stimuli. The threshold for Kullback–Leibler divergence was set to 0.1.

If the regression curve parameters $b^{(-)}$ and $b^{(+)}$ approached 0, the acceptable delay times calculated using equations (2) would become extremely large. However, it would be unnatural for commentary audio, including descriptions of the video, to exhibit such values. Therefore, in this study, the average time interval of utterance start times in the dataset was used as the upper limit for acceptable delay times. For the dataset in this experiment, this value was 5.13 seconds.

2) Participant attributes: Participants were asked to complete a questionnaire on their attributes, as shown in Table I. Tables II show the number of participants for each attribute[3].

TABLE I
Single-choice questionnaire for participant attribute.

ID	Content
01	Age
02	Gender
03	Familiarity with the game (Yes/No)
04	Level of experience with the game
05	Extent of viewing third-person commentary videos of the game
06	Extent of viewing first-person commentary videos of the game

TABLE II
The number of participants of each attribute (excerpt).

ID	Responses				
Q04	Competitive	Enjoy	Sometimes	Only seen	Never
	10	46	305	235	131
Q05	Often	Sometimes	Never		
	34	328	365		

3) Research objectives: This section shows these aspects:

- **Average acceptable delay (Section IV-B)**: measuring the average acceptable delay for commentary audio.
- **Impact of participant attributes (Section IV-C)**: Examining whether acceptable delay varies based on participant attributes listed in Table I.
- **Effects of gameplay content (Section IV-D)**: Examining whether acceptable delay changes based on topic tags.

B. Average acceptable delay

To grasp the overall trends, regression curves and acceptable delay times were determined from all evaluation results. The results are shown in Figure 4. For negative delay times, the regression curve remains almost constant. One possible explanation is the method of setting the reference time. In this experiment, a delay of 0 seconds corresponds to the human commentator's utterance start time. Since the commentary audio itself was delayed relative to the video, applying a negative delay time to such commentary might not have affected its naturalness within a certain range of delays. Addressing this hypothesis would require measuring the delay in the human commentator's audio, which is planned for future work.

The acceptable delay times were determined to be $\mathrm{D50}^{(-)} = 5.13$, $\mathrm{D75}^{(-)} = 5.13$, $\mathrm{D75}^{(+)} = 2.39$, and $\mathrm{D50}^{(+)} = 3.71$ seconds. These values represent average acceptable delays, independent of content and viewer attributes.

C. Impact of viewer attributes

1) Gaming experience: The impact of gaming experience on acceptable delay times was investigated. Figure 5 illustrates the acceptable delay times. It was observed that the deeper the gaming experience, the shorter the acceptable delay time.

This trend aligns with previous studies [16], which suggest that individuals with more experience tend to have faster reaction times to stimuli. For instance, comparing $\mathrm{D50}^{(+)}$, the acceptable delay time for participants with competitive gaming experience ("Competitive") was 2.89 seconds, whereas for those with no experience ("Never"), it was 4.55 seconds—a maximum difference of 1.66 seconds.

2) Game commentary viewing experience: Figure 6 shows acceptable delay times by third-person commentary viewing experience, consistent with trends in Section IV-C1. Participants with deeper viewing experience exhibited shorter

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

[3] https://github.com/takamichi-lab/Video-Stimuli-Super-Smash-Bros

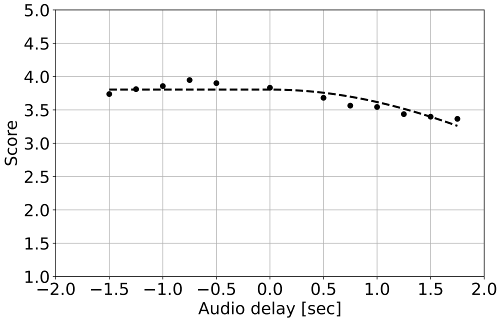

Fig. 4. Average evaluation results and regression curve for all participants.

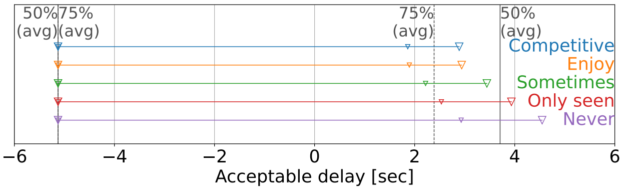

Fig. 5. Acceptable delay times based on gaming experience.

acceptable delay times. For example, participants categorized as “Often” viewers had D75$^{(+)}$ $= 2.18$ seconds, exceeding the overall average by 0.21 seconds. A similar trend was found for first-person commentary experience [4].

This suggests that both gameplay and commentary-viewing experience lead to shorter acceptable delays.

D. Impact of gameplay content

The impact of commentary utterance topics on acceptable delay times was investigated. The analysis method was the same as described in Section IV-C, averaging scores for each utterance topic. Figure 7 shows delay times by topic tag.

The acceptable delay time for the “Result” topic was notably shorter. Specifically, D50$^{(+)}$ $= 3.08$ seconds, approximately 0.63 seconds shorter than the overall average. This can be attributed to the fact that the “Result” topic involves conveying the match outcome at the end of the video, requiring immediate audio commentary once the result is visually displayed. Furthermore, the “Result” topic also exhibited extremely short acceptable delay times for negative delays. This phenomenon likely occurred because applying a negative delay resulted in the commentary audio starting before the result was visually displayed, leading to non-causal commentary.

On the other hand, the “Fighter” topic exhibited significantly longer acceptable delay times compared to other topics. This is likely because the “Fighter” topic falls under color commentary, which does not require immediate delivery.

V. Conclusion

In this study, we evaluated the timing of utterances in game commentary audio in Super Smash Bros. Ultimate. We found a dependency on gaming experience and game contents regarding the acceptable delay.

Acknowledgment: 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.

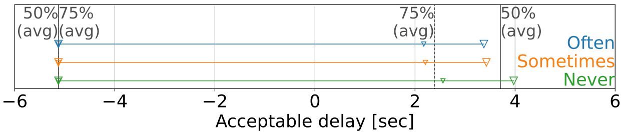

Fig. 6. Acceptable delay times based on third-person commentary viewing experience.

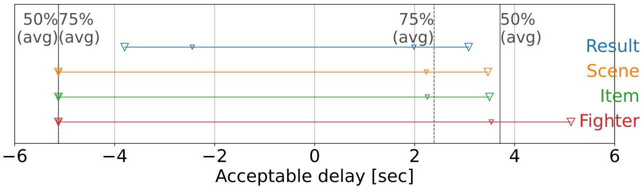

Fig. 7. Acceptable delay times by utterance topic.

References

[1] M. Lee, D. Y. Kim, and P. Pedersen, “Investigating the role of sports commentary: An analysis of media-consumption behavior and programmatic quality and satisfaction,” *Journal of Sports Media*, 2016.

[2] T. Ishigaki, G. Topic, Y. Hamazono, H. Noji, I. Kobayashi, Y. Miyao, and H. Takamura, “Generating racing game commentary from vision, language, and structured data,” in *Proc. INLG*, 2021, pp. 103–113.

[3] J. Rao, H. Wu, C. Liu, Y. Wang, and W. Xie, “MatchTime: Towards automatic soccer game commentary generation,” in *Proc. EMNLP*, 2024.

[4] Y. Taniguchi, Y. Feng, H. Takamura, and M. Okumura, “Generating live soccer-match commentary from play data,” in *Proc. AAAI*, 2019.

[5] K. Shashank Yadav, V. Lohith, M. S. Sachin, M. V. Patil, and S. Narayan, “Automated cricket commentary generation for videos,” in *Proc. CVIP*, 2024, pp. 432–444.

[6] E. Marrese-Taylor, Y. Hamazono, T. Ishigaki, G. Topić, Y. Miyao, I. Kobayashi, and H. Takamura, “Open-domain video commentary generation,” in *Proc. EMNLP*, Y. Goldberg, Z. Kozareva, and Y. Zhang, Eds., 2022, pp. 7326–7339.

[7] K. Zhou, J. Yang, C. C. Loy, and Z. Liu, “Learning to prompt for vision-language models,” *IJCV*, vol. 130, no. 9, 2022.

[8] P. Zhang, X. Li, X. Hu, J. Yang, L. Zhang, L. Wang, Y. Choi, and J. Gao, “Vinvl: Revisiting visual representations in vision-language models,” in *CVPR*, 2021, pp. 5579–5588.

[9] D. H. Klatt, “Review of text-to-speech conversion for English,” *The Journal of the Acoustical Society of America*, vol. 82 3, 1987.

[10] C.-M. Chien, J.-H. Lin, C.-y. Huang, P.-c. Hsu, and H.-y. Lee, “Investigating on incorporating pretrained and learnable speaker representations for multi-speaker multi-style text-to-speech,” in *ICASSP*, 2021, pp. 8588–8592.

[11] M. Armstrong, “The development of a methodology to evaluate the perceived quality of live tv subtitles,” in *IBC2013 Conference Technical Papers*, 2013.

[12] P. Romero-Fresco, “Negotiating quality assessment in media accessibility: the case of live subtitling,” *Univers. Access Inf. Soc.*, vol. 20, 2021.

[13] P. Romero-Fresco and F. Pöchhacker, “Quality assessment in interlingual live subtitling: The NTR Model,” *L. Antv.*, vol. 16, 2017.

[14] A. Watson and A. Fitzhugh, “The method of constant stimuli is inefficent,” *Perception & psychophysics*, vol. 47, pp. 87–91, 1990.

[15] T. Salthouse, “The processing-speed theory of adult age differences in cognition,” *Psychological review*, vol. 103, pp. 403–428, 1996.

[16] B. Repp, “Sensorimotor synchronization: A review of the tapping literature,” *Psychonomic bulletin & review*, vol. 12, pp. 969–992, 2006.

[17] Y. Saito, S. Takamichi, and H. Saruwatari, “SMASH corpus: A spontaneous speech corpus recording third-person audio commentaries on gameplay,” in *Proc. LREC*, 2020, pp. 6571–6577.

[18] R. C. Prim, “Shortest connection networks and some generalizations,” *The Bell System Technical Journal*, vol. 36, no. 6, 1957.

[4] Due to the page limit, we did not show figures in this paper.