Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 6571–6577
Marseille, 11–16 May 2020

SMASH Corpus: A Spontaneous Speech Corpus Recording Third-person Audio Commentaries on Gameplay

Yuki Saito, Shinnosuke Takamichi, Hiroshi Saruwatari
Graduate School of Information Science and Technology, The University of Tokyo, Japan.
7-3-1 Hongo, Bunkyo-ku, Tokyo, Japan.
{yuuki_saito, shinnosuke_takamichi, hiroshi_saruwatari}@ipc.i.u-tokyo.ac.jp

Abstract

Developing a spontaneous speech corpus would be beneficial for spoken language processing and understanding. We present a speech corpus named *the SMASH corpus*, which includes spontaneous speech of two Japanese male commentators that made third-person audio commentaries during the gameplay of a fighting game. Each commentator ad-libbed while watching the gameplay with various topics covering not only explanations of each moment to convey the information on the fight but also comments to entertain listeners. We made transcriptions and topic tags as annotations on the recorded commentaries with our two-step method. We first made automatic and manual transcriptions of the commentaries and then manually annotated the topic tags. This paper describes how we constructed the SMASH corpus and reports some results of the annotations.

Keywords: spontaneous speech corpus, gameplay, third-person audio commentary, transcription, topic tagging

1. Introduction

Spontaneous speech, typically uttered by one person without planning what to say in advance, has many aspects to be studied for deeper understanding of our speech communication. Various techniques for spontaneous speech processing and understanding have been proposed coinciding with the progress in machine learning research (e.g., recognition (Furui, 2003; Toyama et al., 2017), synthesis (Werner et al., 2004; Székely et al., 2019), and summarization (Furui et al., 2004; Szaszák et al., 2017)). Speech corpora including a large variety of spontaneous speech are required to develop these techniques further. Not only the corpora themselves but also methodologies for recording spontaneous speech are important.

The largest spontaneous speech corpus in Japanese is the Corpus of Spontaneous Japanese (CSJ) (Maekawa et al., 2000), which contains about 660 hours of spontaneous speech uttered by 1,417 Japanese speakers with various speaking styles. Various situations for recording diverse spontaneous speech were organized for constructing this enormous corpus (e.g., actual academic presentation, simulated public speaking, and task-oriented dialogues). Although the CSJ is undoubtedly beneficial to both speech and natural language researchers, such a large-scale methodology for the data collection is naturally difficult for other researchers to follow. One easier way to compile naturally uttered spontaneous speech is to make a recording situation where the speakers have many opportunities to see things in their daily lives and to express their thoughts and emotions induced by the situation with no hesitations. The Online Gaming Voice Chat (OGVC) corpus (Arimoto et al., 2012) is one example that utilizes such a situation. A voice chat system was adopted for it as a natural communication tool for speakers to talk with each other while playing well-known massively multiplayer online role-playing games. The OGVC corpus recorded spontaneous and emotional speech through *first-person gameplay*.

In this paper, we present a new methodology for collecting spontaneous speech based on recording *third-person audio commentaries on gameplay*. The core idea behind our methodology is utilizing the audio commentator's ability not only in explaining what is happening during the gameplay (e.g., game states in Japanese chess (Mori et al., 2016)) but also in making some comments that entertain listeners, as shown in Figure 1. For example, when a commentator sees something exciting during the gameplay, he/she is required to explain to the listeners why the action was exciting using his/her expressive speech. Furthermore, the commentator should flexibly and instantly decide on the topic to convey because kaleidoscopic changes occur in the gameplay he/she watches, which can even have some interruptions during the recorded commentaries. Because all we need is to prepare various kinds of gameplay and to hire commentators who are very familiar with the game, our methodology can make a collection of spontaneous speech more easily than ever before. In addition, a collection of pairs of gameplay descriptions and third-person audio commentaries will help us to develop more advanced speech technologies such as multimodal speech synthesis or video summarization involving speech based on visually-grounded machine learning (Roy, 2002).

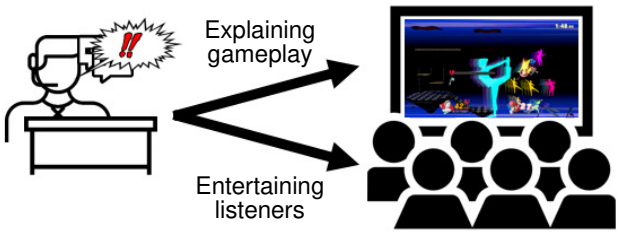


Figure 1: Third-person audio commentary on gameplay. A commentator is required to 1) explain the gameplay scenes and to 2) entertain the listeners by using his/her expressive speech.

We built a spontaneous speech corpus named *the SMASH*

corpus as part of an initial investigation on the effectiveness of our methodology. The corpus includes audio commentaries on the gameplay of the Super Smash Bros. Ultimate (SSBU) [1], a well-known fighting game with easy but exciting content for both players and listeners. We believe that the use of SSBU for our methodology was proper because both players and commentators are required to consider the ad-lib nature of the gameplay such as the sudden appearance of items and the unpredictable moves of the characters on the screen. Our methodology used in this paper consisted of four steps. First, we employed four pairs of SSBU players including four males and four females and recorded their gameplay for about one hour per pair. Second, we hired two Japanese male commentators, who were familiar with SSBU, to record the audio commentaries. Third, we manually split the recorded materials match-by-match. Finally, we annotated transcriptions and topic tags on the recorded commentaries based on our two-step method using automatically generated transcriptions as the initial states before making manual transcriptions.

This paper is organized as follows. Section 2 gives preliminaries of the SSBU gameplay. Section 3 describes how we recorded the gameplay and audio commentaries. Section 4 explains how we annotated the transcriptions and topic tags. Section 5 presents some annotation results. Section 6 concludes this paper with a summary. Section 7 presents future plans. Section 8 describes the availability of the SMASH corpus.

2. Preliminaries of the SSBU Gameplay

SSBU is a well-known fighting game published by Nintendo for the Nintendo Switch. As of October 2019, it had sold over 15.71 million copies worldwide (Nintendo, 2019), earning the title of the best-selling fighting game of all time.

2.1. Basic Framework

Figure 2 shows an example of a scene of the SSBU gameplay involving two players operating Peach and Daisy. The basic framework of the gameplay is quite simple, i.e., each player operate a character called a fighter and tries to damage his/her opponent using the fighter's attacks and force it off the screen, thereby knocking it out (i.e., KOing it). The percentage values shown in the lower part of the screen (e.g., "43.2%" for player 1 and "48.8%" for player 2 in Figure 2) represent the total damage that the fighters have received. The more damage fighters take, the further they are knocked back when hit. Each player scores points by successfully KOing his or her opponent by launching the other fighter off the screen in any direction or by causing the opponent fall off the screen. Basically, the winner is determined by whoever gains the most KOs during the match. Players can also utilize some items randomly generated (e.g., a hammer from the classic Donkey Kong game).

Before starting a match, players choose or randomly select one of various stages and fighters they want to use. Excluding downloadable options, SSBU features 103 stages and 74 fighters. These fighters differ from each other in many aspects such as attack styles, size, damage dealt, kinds of jumps, and movement speed, but how to operate each fighter is basically the same. Because the stages also have different factors (e.g., the size of the stage and features that can harm fighters on it or more easily enable KOs), players are required to grasp the factors and move around the stage adeptly after making their fighter selections.

In the following subsections, we briefly describe some important factors for understanding the SSBU gameplay.

Figure 2: Example of a gameplay scene in SSBU. The screen shows two *fighters* (Peach and Daisy), one guest character (Starman summoned by Peach), and one *item* (Fire Bar) on the *stage* (Dream Land).

2.2. Item Utilization

SSBU features 83 items that make the gameplay more exciting. They can roughly be classified into several categories like shooting, battering, and exploding items, which can be found by playing the "Training" mode of SSBU. We further explain two kinds of items that bring high randomness to the gameplay scenes and significantly affect the gameplay style and audio commentary topics.

2.2.1. The Poké Balls

The Poké Ball (Figure 3(a)) and the Master Ball are thrown by a fighter to randomly summon one of 55 *Pokémons* producing an effect that temporarily assists the player. Although the functions of these two items are the same, the Master Ball releases a Legendary Pokémon that often has a powerful effect to damage the summoner's opponent, as shown in Figure 4.

2.2.2. Assist Trophy

Similar to the Poké Balls, the *Assist Trophy* (Figure 3(b)) randomly releases one of 59 guest characters. Almost all the guest characters support the summoner, but some of them disturb players with their effects such as by darkening the screen, reversing all directional inputs for all players, and turning all fighters invisible. Fighters can KO some of the guest characters to boost their scores when they appear during the stage.

3. Collection of the SMASH Corpus

In this study, we built the SMASH corpus including audio commentaries on the SSBU gameplay. These com-

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

(a) The Poké Ball

(b) The Assist Trophy

Figure 3: The Poké Ball and the Assist Trophy. Using these items, fighters can summon a Pokémon or a guest character that support the player.

Figure 4: Example of a scene after summoning Keldeo, one of the Legendary Pokémons featured in SSBU.

mentaries expressively detail what is transpiring during the gameplay.

3.1. Recording the SSBU Gameplay

For the gameplay recordings, we used only timed matches, with the time limit set to two minutes and thirty seconds. Other rules were set to the default options, and a player who earned the most KOs within the time limit won the game. In the event of a tie, the winner was decided in a sudden-death match. We did not set any limitations for players in the stage and fighter selections. However, no downloadable content was used because we did not download it before the recording period.

For recording the SSBU gameplay, we recruited four pairs of players who had much experience with the game. Specifically, we hired four male players and four female players and covered all gender pairs, i.e., a pair of two males (MM), a pair of two females (FF), and two pairs of males and females (MF1 and MF2), enabling us to record a large variety of gameplay situations. Each pair played SSBU for about an hour offline. During the recordings, we asked each pair to play with two different styles, 1) singles matches with the two players in the first half hour, and 2) doubles matches with two teams consisting of the two players and two non-player character (NPC) fighters in the last half hour. The difficulty levels of the NPC fighters ranged from one (the weakest) to nine (the strongest) and were set to the values the players wanted. Not only the matches, but also the screens of the stage and fighter selections were recorded. The video resolution and frame rate for the recordings were 1920×1080 and 30 fps, respectively.

Table 1: Details of the recorded SSBU gameplay

Player pair	Duration	# of matches Singles	Doubles
MM	60 min 32 sec	9	9
FF	59 min 40 sec	9	8
MF1	58 min 41 sec	9	8
MF2	58 min 18 sec	9	8

Table 1 lists the details of the game. The table shows that the number of doubles matches recorded by the player pair "MM" was nine, which differed from those recorded by the other pairs. One of the reasons was the fact that the pair tended to randomly choose not only the stages but also the fighters, which increased the time to record one match beyond the others. However, the first of the nine doubles matches was different from all the others in terms of its team composition, i.e., they made "two teams of one player and one NPC fighter" for the match. After the match, they created "two teams of two players and two NPC fighters," which was the same as the doubles matches played by the other pairs.

3.2. Recording the Audio Commentaries

We hired two professional Japanese male commentators (MC1 and MC2) who had deep knowledge of SSBU to record the audio commentaries on the gameplay. MC1 made the commentaries on the gameplay pre-recorded by MM, FF, and MF1, and MC2 did the same on the remainder recorded by MF2[2]. Both commentators talked while watching the whole scenes recorded during the gameplay. All of the audio commentaries were recorded in a studio under the administration of a professional sound director, and recording for each commentator was done within one day. All speech samples were recorded using a unidirectional desktop condenser microphone, and the sampling rate was 48 kHz. The 16 bit/sample RIFF WAV was used as the audio format. Figure 5 illustrates the situation for recording the audio commentaries.

4. Annotation of the SMASH Corpus

Making annotations is one of the most difficult tasks in constructing a spontaneous speech corpus because usually we do not prepare any texts to be uttered by the speakers. In this paper, we used a two-step annotation method that utilizes automatically generated transcriptions as the initial states before starting the manual annotations. For the initial investigation, we annotated transcriptions and topic tags on the audio commentaries that we recorded.

4.1. Data Preprocessing

Although the recorded gameplay consisted of many scenes including the stage and fighter selections and the singles or

[2]In our methodology, we can simultaneously record the gameplay and audio commentaries through a running commentary on live gameplay (e.g., official tournaments).

Figure 5: Recording audio commentaries on the SSBU gameplay. We put a desktop microphone, and the commentator gave an audio commentary while watching the pre-recorded gameplay.

doubles matches, we focused on the matches parts as the subjects to be annotated because they were the main contents of the SSBU gameplay. Therefore, we divided each of the recorded matches and corresponding audio commentaries into eight or nine parts match-by-match. Each of the divided gameplay contained 1) the fighter introduction part (about three seconds), 2) the match part (about two minutes and forty five seconds, including a sudden-death match if it occurred), and 3) the results part (about ten seconds).

4.2. Automatic Transcriptions

Making transcriptions of speech from scratch should be hard, especially in the case of spontaneous speech. Therefore, we made use of the recently-developed cloud-based speech-to-text (STT) service for automatically generating transcriptions that would help annotators do their tasks. We utilized the Google Cloud STT[3] and generated an automatic transcription for each of the divided audio commentaries.

4.3. Manual Annotations

On the basis of the automatic transcriptions, we made 1) refined transcriptions and 2) topic tags as annotations to the audio commentaries. We employed Japanese annotators (without hearing impairment) who did these two tasks simultaneously while watching the gameplay videos mixed with their corresponding audio commentaries.

Before the annotations, we manually split each of the commentaries into several segments considering segment duration, sentence length, and breath insertion using the automatic transcriptions. We recorded the start and end times of each segment in the original audio commentary and provided the interval information to the annotators for their convenience. Finally, 2,085 segments (i.e., approximately 30 segments per match) of the commentaries were obtained.

4.3.1. Refinement of the Automatic Transcriptions

The annotators refined the automatic transcriptions while watching the gameplay videos with audio commentaries. For the refinements, we asked the annotators not to insert any punctuations because defining the duration of the punctuation in spontaneous speech was a difficult task. We also instructed them not to transcribe fillers and falterings that were completely missing in the automatic transcriptions. Proper nouns that the annotators transcribed but did not feel confident about in the results were firstly written in katakana and then manually modified by the first author of this paper. Finally, 1.1 million words were included in the refined transcriptions[4].

4.3.2. Topic Tagging to the Audio Commentary

The participants also annotated a topic tag for each segment of the audio commentaries considering both the commentaries and corresponding gameplay scenes. In this work, we defined the topic tags listed in Table 2 taking important factors of the SSBU gameplay into account.

Table 2: List of topic tags defined in this work

Tag	Definition
Fighter	A fighter on the screen, or a series that featured the fighter
Stage	A stage on the screen, or a series that featured the stage
Item	An item on the screen, or a series that featured the item
Pokémon	A Pokémon on the screen
Assist Trophy	An Assist Trophy on the screen
Match	A scene on the screen, except for the aforementioned content
Result	A result of the match
Chat	Anything but the scene on the screen, like a commentator's past experience

5. Annotation Results

5.1. Visually Grounded Annotations

Table 3 lists some examples of the annotation results. This table shows that some of the refined transcriptions included a key word (e.g., the names of the game) that characterized the topic to be discussed. Figure 6 illustrates the gameplay scenes corresponding to the annotation results in Table 3.

5.2. Topic Transition

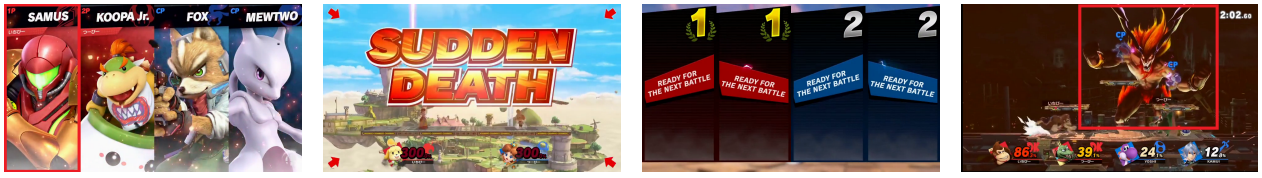
Figure 7 shows an example of a topic transition in audio commentaries on a singles match. This figure illustrates that the commentaries covered all of the topics listed in Table 2, and there were some sudden transitions (i.e., interruption) caused by unpredictable moves of characters on the screen. For example, a Final Smash, i.e., incredibly powerful attacks of a fighter, changed the topic from "Chat" to "Match" suddenly. We observed that the main topics of the overall commentaries were either "Fighter" or "Match," and some random factors such as "Item," "Pokémon," and "Assist Trophy" tended to cause the interruption of the commentaries.

[3] https://cloud.google.com/speech-to-text/

[4] Here, we used MeCab (Kudo, 2006) as a morphological parser and NEologd (Sato, 2015) as a dictionary for counting the number of words.

Table 3: Examples of annotation results. The second column lists transcriptions translated into English. The underlined text in the second column denotes a proper noun (with its brief explanation in English text). We transcribed phonemes in the third column for reference

ID	Refined transcription with translation in English	Phonemes	Topic tag
(1)	もうねメトロイドのエンディングでね中身が女性だっていうことが分かるんですけれども In the ending of Metroid (name of the game), we discovered that the main character was female.	m o u n e m e t o r o i d o n o e N d i N g u d e n e n a k a m i g a j o s e i d a q t e i u k o t o g a w a k a r u N d e s u k e r e d o m o	Fighter
(2)	さあサドンデスなりましたどっちが勝つのか Now a sudden-death match starts, which player win the match?	s a a s a d o N d e s u n a r i m a s i t a d q ch i g a k a ts u n o k a	Match
(3)	じゃあもうこの二人はもうレベル9で行った方がいいんじゃないかなと思いますけれどもね I believe that the team of the two players can beat NPC fighters with their difficulty levels set to nine.	j a a m o u k o n o f u t a r i w a m o u r e b e r u k y u u d e i q t a h o u g a i i N j a n a i k a n a t o o m o i m a s u k e r e d o m o n e	Chat
(4)	さあおおイフリート出ましたね Now, oh, Ifrit (name of the character) appears in the stage.	s a a o o i f u r i i t o d e m a sh i t a n e	Stage



(1) Fighter introduction (2) Sudden-death match (3) Result (4) Doubles match

Figure 6: Examples of gameplay scenes corresponding to the annotation results listed in Table 3. In (1) and (4), the commentator was talking about the red-boxed objects, i.e., (1) Samus (the main character of Metroid) and (4) Ifrit (a character appearing in the stage).

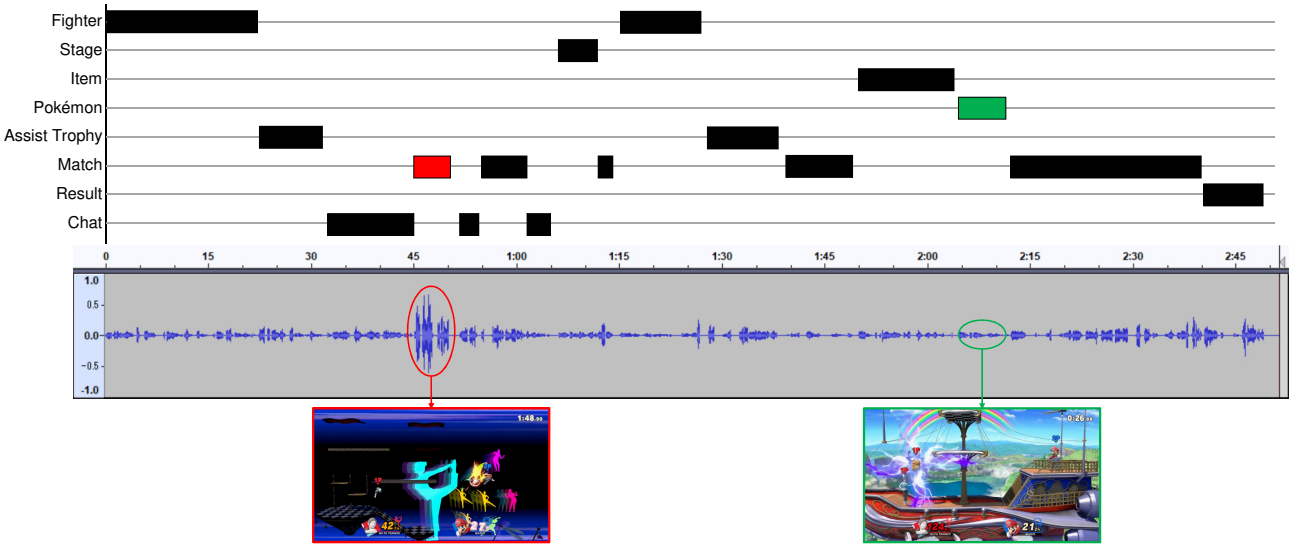


Figure 7: Example of topic transition in audio commentaries. The topic was changed by suddenly occurred events, e.g., Final Smash (incredibly powerful attacks) marked with a red circle and a Pokémon marked with a green circle. Note that successive segments having the same topic tags were merged for clear illustration.

5.3. Effectiveness of Automatic Transcriptions

To investigate the effectiveness of our two-step transcription method, we calculated the word error rate (WER) of the automatic transcriptions using the refined ones as the reference and found that the WER averaged across annotated speech segments was about 10.3%. This indicates that using the automatic transcription can reduce the annotators' loads compared with doing annotation from scratch.

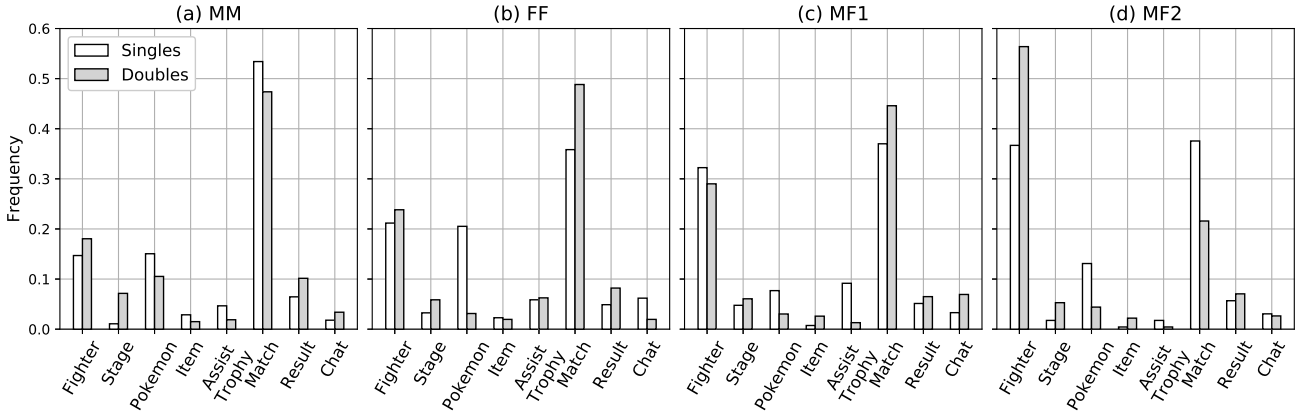


Figure 8: Distribution of annotated topic tags. Note that the commentator for the gameplay videos recorded by (d) MF2 was different from the others.

5.4. Distributions of Topic Tags

Figure 8 shows the distributions of the topic tags. Here, the annotation results were summarized by players in a pair wise manner. This figure shows that the three distributions of the topic tags shown in Figure 8(a)–(c), where commentator MC1 made the audio commentaries, have similar shapes, i.e., the most frequent tags are "Match" regardless of the player pairs. However, the distribution shown in Figure 8(d) is significantly different from the others. These results suggest that 1) the same commentator tended to make similar audio commentaries even if the players changed, as long as the settings of the gameplay remained unchanged, and 2) the difference in the commentators greatly affected the topics of the commentaries in accordance with their personalities and preferences.

6. Conclusion

We built a spontaneous Japanese speech corpus named the SMASH corpus using our methodology that records third-person audio commentaries on gameplay. The corpus contains about four hours of gameplay from Super Smash Bros. Ultimate and corresponding audio commentaries made by two Japanese males. We annotated 1) automatic and manual transcriptions and 2) topic tags for the recorded commentaries with our two-step annotation method using automatic transcriptions as the initial states before starting annotations. We presented some results of the annotations, and found that 1) the automatic transcriptions could be a good initialization for manual refinement and 2) the difference in the commentators significantly affected the distributions of the annotated topic tags.

7. Future Directions

7.1. More Complicated Settings of Gameplay

Although we used the most basic style of the SSBU gameplay that had only two players, we can involve up to eight players in one match offline or online to make more exciting situations to be recorded. Also, we can employ some professional players for recording more sophisticated gameplay scenes. We will continue to record more matches to investigate the effects caused by changes in the gameplay settings.

7.2. Other Style and Language of Commentaries

In our work, two commentators (MC1 and MC2) watched different matches in recording the commentaries. We can also show the same match to different commentators to investigate the variation observed in their audio commentaries. As with the recording of the Crowdsourced Parallel Japanese Dialects (CPJD) corpus (Takamichi and Saruwatari, 2018), we can easily recruit different commentators with the use of a web-based recording platform and a crowdsourcing service. Moreover, we can hire commentators other than Japanese native speakers because SSBU is played worldwide and it supports 11 languages including English, Chinese, French, and Spanish.

7.3. Multimodal Annotations

We annotated the topic tag to each segment in this paper. However, as we focused on the audio commentary regarding gameplay drastically changing over time, topics within each segment can also be changed. We will annotate time-varying topic tags to the audio commentaries with the use of multimodal annotation tools such as Anvil (Kipp, 2001). Also, we will made more useful annotations like emotion induced by the gameplay scenes.

7.4. Automatic Audio Commentary Generation

As an example of the visually-grounded speech technologies mentioned in the third paragraph of Section 1 , we will develop the automatic audio commentary system using the SMASH corpus. This system involves many challenging tasks, e.g., audio-visual scene analysis (Owens and Efros, 2018) and sentiment analysis of a commentator, context management (Jokinen et al., 1998), and incremental text-to-speech synthesis (Baumann, 2014). We will investigate a way to realize such system and to evaluate the system itself from various viewpoints like naturalness, validity, and interest of the commentaries.

8. Availability of the SMASH Corpus

Although the current version of the SMASH corpus is publicly unavailable, we are doing the preparation to make the corpus available online for non-commercial research. One

can contact with the first author to get the SMASH corpus for his/her own study.

9. Acknowledgements

Part of this work was supported by JSPS KAKENHI Grant Number 18J22090 and 19H01116. The authors thank Naoko Tanji, Hiroki Tamaru, Takaaki Saeki, and Kentaro Mitsui, members of the System #1 Lab. at the University of Tokyo, for their insightful comments.

10. Bibliographical References

Arimoto, Y., Kawatsu, H., Ohno, S., and Iida, H. (2012). Naturalistic emotional speech collection paradigm with online game and its psychological and acoustical assessment. *Acoustical Science and Technology*, 33(6):359–369, June.

Baumann, T. (2014). Partial representations improve the prosody of incremental speech synthesis. In *Proc. Interspeech*, pages 2932–2936, Singapore, Sep.

Furui, S., Kikuchi, T., Shinnaka, Y., and Hori, C. (2004). Speech-to-text and speech-to-speech summarization of spontaneous speech. *IEEE Trans. on Speech and Audio Processing*, 12(4):401–408, July.

Furui, S. (2003). Recent advances in spontaneous speech recognition and understanding. In *Proc. SSPR*, Tokyo, Japan, Apr.

Jokinen, K., Tanaka, H., and Yokoo, A. (1998). Context management with topics for spoken dialogue systems. In *Proc. COLING*, pages 631–637, Montreal, Canada, Aug.

Kipp, M. (2001). Anvil - a generic annotation tool for multimodal dialogue. In *Proc. Eurospeech*, pages 1367–1370, Aalborg, Denmark, Sep.

Kudo, T. (2006). Mecab: Yet another part-of-speech and morphological analyzer. `http://taku910.github.io/mecab`. visited on Mar. 5, 2020.

Maekawa, K., Koiso, H., Furui, S., and Isahara, H. (2000). Spontaneous speech corpus of Japanese. In *Proc. LREC*, pages 947–952, Athens, Greece, May.

Mori, S., Richardson, J., Ushiku, A., Sasada, T., Kameko, H., and Tsuruoka, Y. (2016). A Japanese chess commentary corpus. In *Proc. LREC*, pages 1415–1420, Portorož, Slovenia, May.

Nintendo. (2019). Top selling software sales units - Nintendo Switch software. `https://www.nintendo.co.jp/ir/en/finance/software/index.html`. visited on Dec. 2, 2019.

Owens, A. and Efros, A. A. (2018). Audio-visual scene analysis with self-supervised multisensory features. In *Proc. ECCV*, Munich, Germany, Sep.

Roy, D. K. (2002). Learning visually-grounded words and syntax for a scene description task. *Computer Speech and Language*, 16(3–4):353–385, July.

Sato, T. (2015). Neologism dictionary based on the language resources on the web for mecab. `https://github.com/neologd/mecab-ipadic-neologd`. visited on Mar. 5, 2020.

Szaszák, G., Tündik, M. A., and Beke, A. (2017). Summarization of spontaneous speech using automatic speech recognition and a speech prosody based tokenizer. In *Proc. IC3K*, pages 221–227, Funchal, Portugal, Nov.

Székely, E., Henter, G. E., Beskow, J., and Gustafson, J. (2019). Spontaneous conversational speech synthesis from found data. In *Proc. Interspeech*, pages 4435–4439, Graz, Austria, Sep.

Takamichi, S. and Saruwatari, H. (2018). CPJD corpus: Crowdsourced parallel speech corpus of Japanese dialects. In *Proc. LREC*, pages 434–437, Miyazaki, Japan, May.

Toyama, S., Saito, D., and Minematsu, N. (2017). Use of global and acoustic features associated with contextual factors to adapt language models for spontaneous speech recognition. In *Proc. Interspeech*, pages 543–547, Stockholm, Sweden, Aug.

Werner, S., Eichner, M., Wolff, M., and Hoffmann, R. (2004). Toward spontaneous speech synthesis-utilizing language model information in TTS. *IEEE Trans. on Speech and Audio Processing*, 12(4):436–445, July.