IBC | 2021

TV-WATCHING ROBOT: TOWARD ENRICHING MEDIA EXPERIENCE AND ACTIVATING HUMAN COMMUNICATION

Y. Hagio, M. Kamimura, Y. Hoshi, Y. Kaneko, and M. Yamamoto

NHK, Japan

ABSTRACT

Viewing TV with family and friends makes the experience more enjoyable. Recently, as more people watch TV programs on their mobile devices via the Internet, opportunities to watch TV in groups are decreasing. We propose a companion robot to enhance human communication during the TV-viewing experience. The robot extracts keywords from the video, audio, and subtitle data of the TV program being watched, generates utterances from the keywords, and asks people questions related to the TV program. Viewers can chat with the robot, triggered by its questions. To confirm the utility of the robot, we conducted an experiment in which users watched TV with the robot. Results indicate that over 70% of the participants responded that the robot "promoted active conversations among people” and "created a more relaxed atmosphere." Thus, our study demonstrates that a TV-watching robot can create novel and rich media experiences and stimulate communication among people.

INTRODUCTION

Televisions (TVs) are installed in the living rooms of most homes and are a part of daily life. Watching TV not only provides information or entertainment, but also provides an opportunity for individuals to have conversations and share feelings of joy, anger, and sadness that are triggered by topics related to the TV programs being watched. Furthermore, it serves as a tool of communication between generations of parents, children, and grandchildren. In recent years, more people are using mobile devices to enjoy video content via the Internet, and the opportunities for multiple people to watch TV together in the same room are decreasing. In addition, the average number of people per household in many countries, including Japan, is decreasing (1), making it impossible for many people to watch TV socially, even if they wanted to.

Social robots have become more widespread in recent years with the development of machine learning technology. In the future, it is expected that robots will become even more involved in our lives. The use of social robots in the TV-viewing environment has the potential to increase conversation, provide a relaxed atmosphere, and create a new media experience that is different from the usual TV-viewing experience. It is also expected that the robot's speech will have the effect of increasing people’s attention toward their TV. To verify whether the presence of a robot in a TV-viewing environment can generate these effects, we need a robot that can operate autonomously using TV programs and the people around it as information sources.

Based on prior research, we are developing a social robot that uses TV programs as a source of information for utterances and gestures. Hoshi et al. (2) conducted a dialogue analysis that classified the types of conversations that people have among themselves when watching TV, as a mode of assessing their behavior. They found that the “disclosure” utterances, wherein individuals stated their feelings or thoughts, tended to be the most common forms of spontaneous utterances. They found that their partner’s response rate tended to be high for “question,” “edification,” and “confirmation” utterances.

In this study, we thoroughly describe the functional requirements of a robot that watches TV with people and can communicate with those people based on the results of the dialogue analysis (2). A novel prototype robot was developed based on these functional requirements. The experiments were conducted with pairs of individuals watching TV.

RELATED WORK

In this section, we introduce the work related to robots and interactive agents that operate in TV-viewing environments.

Ogawa et al. (3) proposed a method for operating Internet of Things (IoT) devices, including robots, by distributing metadata in tandem with TV programs using Hybridcast. The advantage of this approach is that the robots can be operated based on the TV program being watched. However, the cost for broadcasters for producing TV programs could increase due to the cost of creating the metadata for operating robots. In addition, because the broadcaster distributes the robot's utterances and other information as metadata, all robots in each household would have the same utterances. Therefore, this method would fail to increase empathy between people and robots in the TV-viewing environment, which is our main goal.

Goto et al. (4) proposed a dialogue agent system that searched for TV programs through natural language dialogue. They searched for programs using natural language based on aspects of the programs like titles and genres. This method enabled users to use natural language to interactively perform operations that would be conventionally done by remote control. However, the method of Goto et al. (4) was a proposal to enhance the functionality of TV receivers, solely for improving convenience, and therefore, is different in direction from our goal of reproducing and enhancing TV viewing by multiple people.

Nishimura et al. (5) proposed a robot that used comments posted on social media (e.g., Twitter) to chat with users. This approach does not require the broadcaster to create metadata to be used for operating the robots. However, when users were watching a TV program with low ratings, the number of robot utterances decreased because there were fewer comments on social media about that program, on which to base its utterances.

Thus, there have been various attempts to introduce robots and interactive agents into the TV-viewing environment. However, most of them have been aimed at supplementing the functions that lack in TV contents or receivers and at enhancing convenience. Therefore, they do not aim to increase or enhance communication during TV viewing using robots, as is our goal. Additionally, the chatting robot proposed by Nishimura et al. (5) used sentences posted on social networks for speech, and thus, no robot has been proposed that generates utterances directly from the TV content itself.

FUNCTIONAL REQUIREMENTS

Our proposed robot needs to act as a partner who enjoys watching TV with people; thus, the robot should behave in a manner similar to that of people watching TV. Hoshi et al. (2) conducted a dialogue analysis that classified the types of utterances between humans when watching TV to determine their behavior while watching TV. A summary of their results, which focused on the utterance behavior of humans while watching TV, is as follows:

- The “disclosure” utterance, wherein individuals state their feelings or thoughts, was the most common form of utterance triggered by the TV program, comprising approximately 33% of all the utterances.
- The response rate of the partner was high for “question,” “edification,” and “confirmation” utterances. The highest response rate was for “question” utterances, which explicitly request a response from the partner, with a response rate of approximately 93%.
- The length of the utterances was 20 Japanese characters or less for 79% of the utterances.

In addition, non-verbal information such as gestures, eye contact, and facial expressions play an important role in daily conversations, including those during TV viewing. Furthermore, with respect to robots, blinking LEDs and vibrations could be used as a means of communicating similar information to these non-verbal gestures. In a TV-viewing environment, it is considered important to use such non-verbal information to express the robot's feelings in a way that does not interfere with the TV viewing of the people watching with the robot.

Based on our findings, we designed a robot to watch TV with people, with technical developments focused on the following functions:

- The gestures of the robot indicate enjoyment while watching TV in a manner similar to that of humans.
- The robot makes “disclosure” utterances, which indicate its feelings or thoughts regarding the TV program being watched.
- The robot makes utterances corresponding to “questions,” “edifications,” and “confirmations” associated with the TV program being watched and starts a conversation based on this.
- The robot accurately responds to the “question,” “edification,” and “confirmation” utterances from the people with whom it is watching.

As a first step toward achieving a robot with the above capacities, we prototyped a robot that actively operates and interacts with humans. We installed the robot with the three functions listed below, after which we conducted an experiment in which people and the robot watched TV together, to verify the robot’s functionality.

Function 1: The robot voluntarily performs physical gestures as if enjoying watching TV or conversing with people.

Function 2: The robot makes utterances relating to its feelings or thoughts concerning the TV program being watched (i.e., disclosure utterance) while facing the TV.

Function 3: The robot asks questions relating to the TV program being watched while facing the people (i.e., question utterance) and converses with the people based on this.

HARDWARE CONFIGURATIONS

We used CommU (6) for implementation of the robot, as shown in Figure 1. CommU is equipped with four motors in both arms, two in the body, three in the neck, three in the eyes, and one in the mouth, allowing the robot to express facial expressions while moving its body. In addition, CommU has a wide range of motion in the horizontal, rotational plane and can turn its head toward people in any direction around the robot.

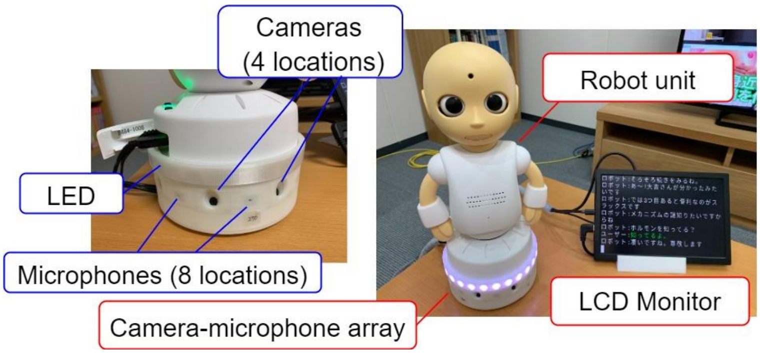


Figure 1 – Hardware configuration of our robot

We prototyped a camera-microphone array, which has a total of four cameras and eight microphones, and installed it as an additional device at the bottom of the CommU. The robot can detect people nearby by using the four cameras installed in the camera-microphone array, within a range of approximately 200 degrees. Furthermore, the robot can distinguish the difference between a human voice and TV sound using sound localization and separation through the open-source audition software HARK (7) and using the eight microphones installed in the camera-microphone array. The detected human voice is then converted into text by speech recognition, using the Speech Service in Microsoft Azure (8), for use in conversation. In addition, the results of the human sound localization are used to determine the positions of the humans.

A band-shaped LED, which can be seen from all directions, is also installed in the camera-microphone array. This LED is used to notify the user that the microphone is turned on.

The people who are watching TV with the robot may not be able to hear what the robots are saying because they are often focused on the TV. Therefore, an LCD monitor, which shows the utterances of the robot and the results of the speech recognition, is placed next to the robot. In addition, the people who are interacting with the robot can confirm the speech recognition results of their utterances.

SOFTWARE DESIGN

The state transitions and the robot's operations at each state for our prototype are shown in Figure 2. The prototyped robot has the three following states: "TV-watching," "disclosure

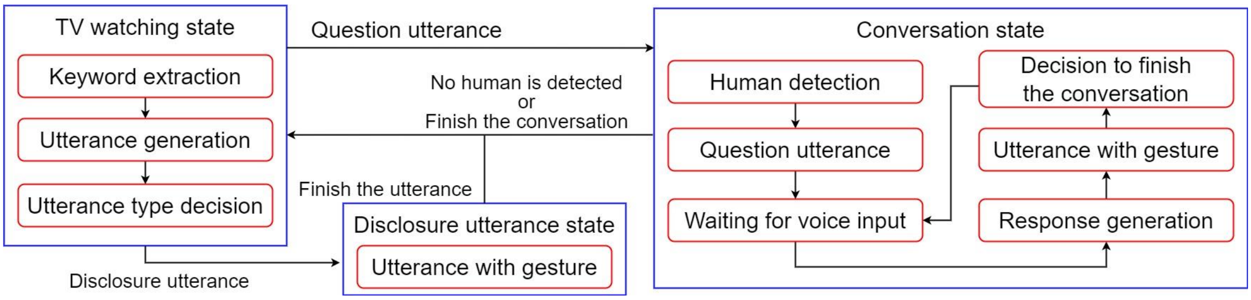


Figure 2 – State transitions and the operations at each state for our robot

utterance," and "conversation." Each state is explained below. Note that although this study presents each function in English, the robot functions in Japanese.

TV-Watching State

In the TV-watching state, the robot extracts keywords from the TV program being watched and generates disclosure and question utterances from those keywords. Then, it randomly decides whether to use a disclosure or question utterance and transitions to the respective state. Furthermore, parallel to this processing, the robot faces the TV and ceaselessly repeats in a random order the following operations, as though it is watching TV: "blinking," "nodding," "moving the neck," and "moving the upper body."

Keyword extraction

The robot obtains the video, audio, and caption data from the TV program being watched, and from that multimedia data it extracts the keywords to be used for utterance generation. Note that our robot has functions that are designed specifically to be used for Japanese TV broadcasting.

For the caption data, the robot extracts the keywords using a keyword dictionary consisting of approximately 160,000 words. The sentences in the caption data are converted into word-separated writing using MeCab (9) with mecab-ipadic-NEologd (10), and the words included in the keyword dictionary are extracted as the keywords.

Some TV programs in Japan have caption data, but others do not. Therefore, for broader applicability, it is necessary to develop a keyword extraction method that does not depend on the caption data. To this end, keywords are also extracted from images and sounds of the TV program.

Keyword extraction based on a keyword dictionary is conducted in a manner similar to that for caption data, after the audio of the TV program has been converted to text using speech recognition via a cloud service. Meanwhile, keywords are extracted through object detection on a local computer. The robot uses Faster-RCNN (11), which was trained using the Microsoft COCO database (12), for object detection on that same computer. The classification results of the detected objects are extracted as keywords. In addition, saliency estimation is conducted for the TV images. To extract suitable keywords from TV images, we take the results of the object classifications whose mean saliency estimate values (13) in rectangular regions of interest obtained from the object detection system and compare them against a threshold value. Words that exceed this threshold are extracted as a keyword.

An example of keyword extraction using object detection on a local computer is shown in Figure 3. The green rectangle indicates the detected objects and the red area indicate high saliency in the estimation results. In this example, the following two objects are detected: "elephant" and "person." However, the "person" has low saliency estimation results; therefore, only the "elephant" is extracted as

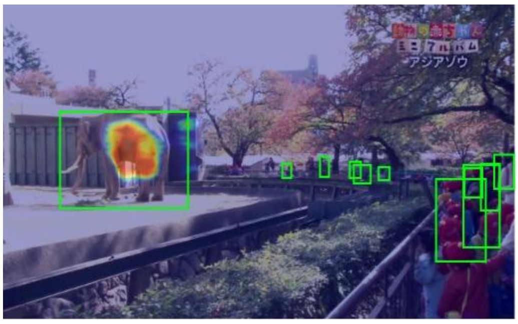


Figure 3 – Example of keyword extraction using object detection on a local computer

the keyword.

The robot also extracts keywords using OCR (Optical Character Recognition), celebrity recognition, descriptions of the content, object detection, and tags extracted from cloud services. We used Microsoft Azure (8) and Amazon Web Service (14) for the cloud services.

Utterance generation

The robot's utterances are generated using the extracted keywords. At this stage of development, the following two types of sentences are generated: "disclosure utterances," which express the robot's feelings; and "question utterances," which ask questions to the surrounding people.

We used the captions from the TV programs on a total of 10 channels from the past seven years as template sentences for the disclosure utterance generation. Our method of utterance generation is shown in Figure 4. The generated disclosure utterances should include the robot's feelings; therefore, we used sentences that express emotion (e.g., "I want to eat," "I want to go") as template sentences.

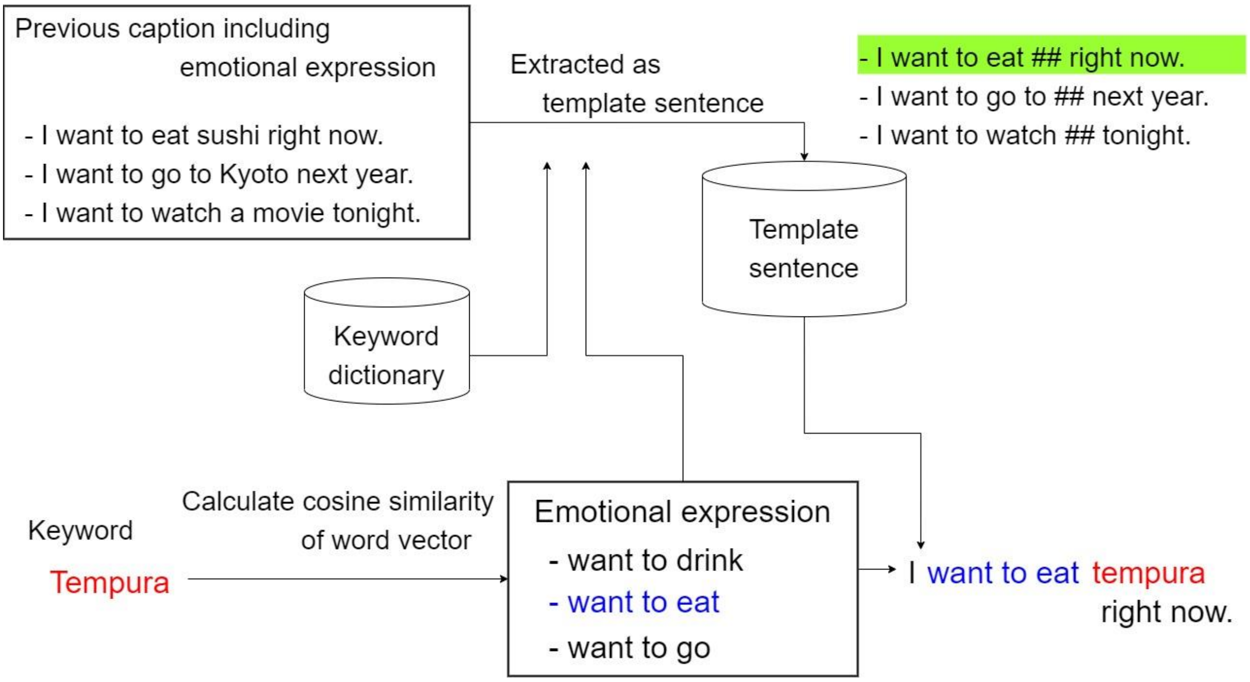

Previous caption including emotional expression
- I want to eat sushi right now.
- I want to go to Kyoto next year.
- I want to watch a movie tonight.
Extracted as template sentence
- I want to eat ## right now.
- I want to go to ## next year.
- I want to watch ## tonight.
Template sentence
Keyword dictionary
Keyword
Tempura
Calculate cosine similarity of word vector
Emotional expression
- want to drink
- want to eat
- want to go
I want to eat tempura right now.

Figure 4 – Example of utterance generation. The red text indicates keywords, and the blue text indicates emotional expressions suitable for the keywords, and the green marker indicates the selected template sentence

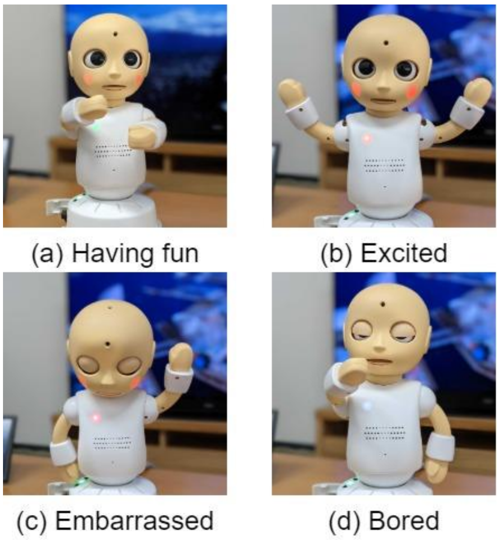
The words included in the keyword dictionary that are used to extract keywords are converted to blanks (expressed as ## in Figure 4) from prior captions which included emotional expressions that were extracted as template sentences. We used captions that had less than 20 characters in Japanese because prior research has shown that 79% of the utterances between humans while watching TV are less than 20 characters in Japanese (2). In addition, past captions are used to learn the distributed representation of words using the Word2Vec technique (15). The extraction of the template sentences and the learning of the distributed representation of the words are conducted in advance of the TV-viewing event. During an interaction, the robot determines the word to use for emotional expressions by calculating the word with the highest cosine similarity to the extracted keywords. A disclosure utterance, which includes an emotional expression, is then generated by inserting the keyword into the template sentence that has the word to use for emotional expression. Keywords used in utterance generation will not be reused until the time set in advance by parameters has elapsed to ensure diversity and variability in the types of utterances.

Next, the robot randomly determines whether to use the generated disclosure or question utterances. The proportion of disclosure and question speeches are determined in advance as parameters, and the robot selects one of these utterance types based on this proportion.

The utterance interval can also be adjusted to avoid situations where the robot is constantly speaking. The utterance frequency is set in advance, and the utterance interval is determined using a Poisson distribution. The robot transitions into a disclosure utterance state when a disclosure utterance is selected, or a conversation state when a question utterance is selected.

Disclosure Utterance State

In the disclosure speech state, the robot faces the TV and speaks the generated disclosure utterances. For text-to-speech synthesis, we used the library that comes standard with CommU (6). The robot simultaneously makes physical gestures when speaking. We made ten physical gestures (e.g., “fun,” “excited,” “boring”) for CommU in advance, from which one is randomly selected and executed. Examples of physical gestures are shown in Figure 5. The selected utterance is displayed on the LCD monitor while the robot is speaking. Once the utterance has been completed, the robot’s current state transitions to the TV-watching state.



(a) Having fun

(b) Excited

(c) Embarrassed

(d) Bored

Figure 5 – Examples of physical gestures

Conversation State

In the conversation state, the robot conducts a conversation triggered by a generated question utterance. The robot detects nearby humans using the camera array and questions them. If multiple humans are detected as estimated by the human detection system, the robot then randomly selects one to face and turns to the appropriate angle. After asking the question, if the human’s response is obtained from the microphone array, then the next utterance is generated from the results of the user’s speech recognition. The robot identifies and speaks to the user who replied to the robot using the result of sound localization, incorporating use of its physical gestures. After finishing a series of back-and-forth responses comprising the conversation, the robot turns toward the TV and its current state transitions to the TV-watching state.

In our prototype robot, we implemented a dialogue function using a dialogue engine (16). This dialogue engine calculates the degree of breakdown in the dialogue between the user and the robot and if the threshold is exceeded, the topic can be changed to allow for continuous dialogue. This allows for the dialogue to not only include things related to the keywords directly obtained from TV programs, but also dialogue on a wide range of topics derived from those keywords.

EVALUATION

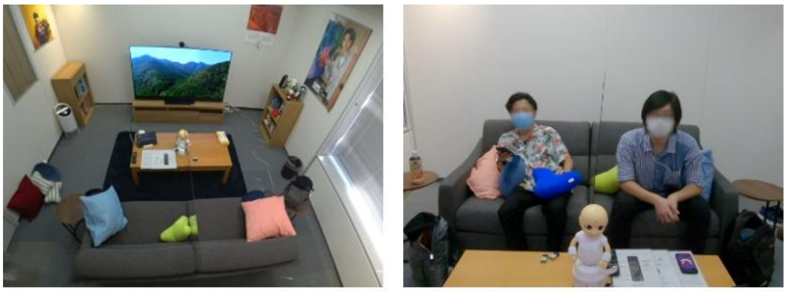
To verify the effectiveness of our prototype robot, we conducted an experiment in which we watched TV with the robot.

Outline of the Experiment

The participants in the experiment were 16 pairs of close friends (32 people in total) who liked to watch TV. The participants in the experiment were men and women, ranging in age from being in their 20s to 70s. The duration of the experiment was 4h for each group, and questionnaires were administered at the beginning and end of the experiment. To ensure that the results of the experiment did not depend on the interest in the programs being watched, the participants could watch their favorite programs from among 10 TV channels selected from the past 7 years of recorded programs. In addition, for the participants to be able to participate in the experiment in the same relaxed state as when watching TV at home, they could take breaks freely during the experiment, and could watch TV while eating, drinking, using a smartphone, reading, or doing other things.

Informed consent was provided before the start of the experiment, and the participants' willingness to participate in the experiment was confirmed after explaining that the experiment could be freely suspended at any time at the will of the participants. In addition, this experiment was approved by the research ethics board of NHK Science & Technology Research Laboratories.

The prototyped robot was placed on a table, as shown in Figure 6 (a), and it operated independently without experimenter control. The disclosure and question utterance for the robot was set to a proportion of 3:1, and the utterance frequency was set to once every 80 seconds on average. The actual conditions of the experiment are shown in Figure 6 (b).



(a) The experimental room (b) The actual condition of the experiment

Figure 6 – Condition of the experimental room and the experiments

Experimental Results

The questionnaire survey items are shown in Table 1. There were nine questions, consisting of four questions on the TV-viewing experience with the robot (Q1-1 – Q1-4), three questions on the robot's behavior (Q2-1 – Q2-3), and two questions on the experimental environment (Q3-1 and Q3-2). Q1 and Q3 were to be answered using a 7-point Likert scale, and Q2 included multiple-choice questions. For Q3, we also asked about the reasons for the answers.

Items	Questions
Q1-1	Has the presence of the robot increased conversation between the two of you?
Q1-2	Did the presence of the robot relax the atmosphere?
Q1-3	Have you turned your attention to the TV as a response to the robot's utterances?
Q1-4	Did the robot's presence make this experience more interesting than your usual TV viewing?
Q2-1	What did you think of the robot's disclosure utterances?
Q2-2	What did you think of the conversation with the robot?
Q2-3	What did you think of the robot's movements and gestures?
Q3-1	Did you enjoy the experiment?
Q3-2	Did you feel tired from the experiment?

Table 1 – Questionnaire survey items

About the robot

First, we examined the effects of watching TV with a robot (Q1). We examined the following four effects that can be expected from watching TV with a robot: increasing conversation between people (Q1-1), making the atmosphere more relaxed (Q1-2), turning to the TV as a response to the robot's utterances (Q1-3), and finding the experience more interesting than usual TV viewing (Q1-4). The results of the survey are shown in Figure 7. The seven levels of responses used in the Likert responses for Q1 and Q3 ranged from strongly agree (7) to strongly disagree (1).

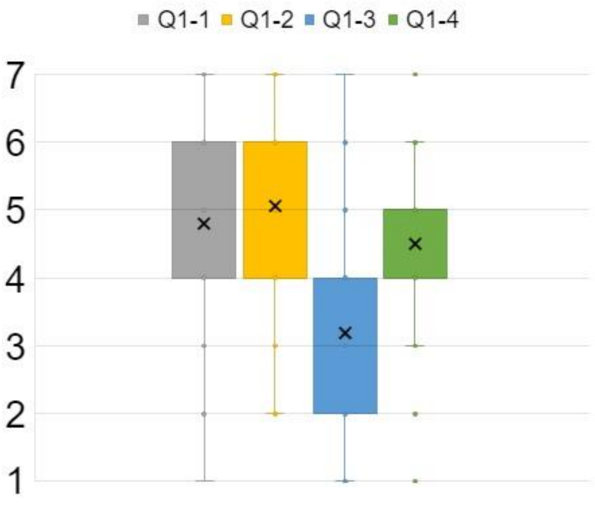


Figure 7 – Results of the questionnaire on TV viewing with robots

The results for the increase in conversation between the two people due to the presence of the robot (Q1-1) were relatively highly rated, with a mean score of 4.8 and a standard deviation of 1.4. Approximately 69% of the participants responded that the presence of the robot increased their conversation, and we confirmed that they recognized the presence of the robot in the TV-viewing environment as what increased their rate of communication.

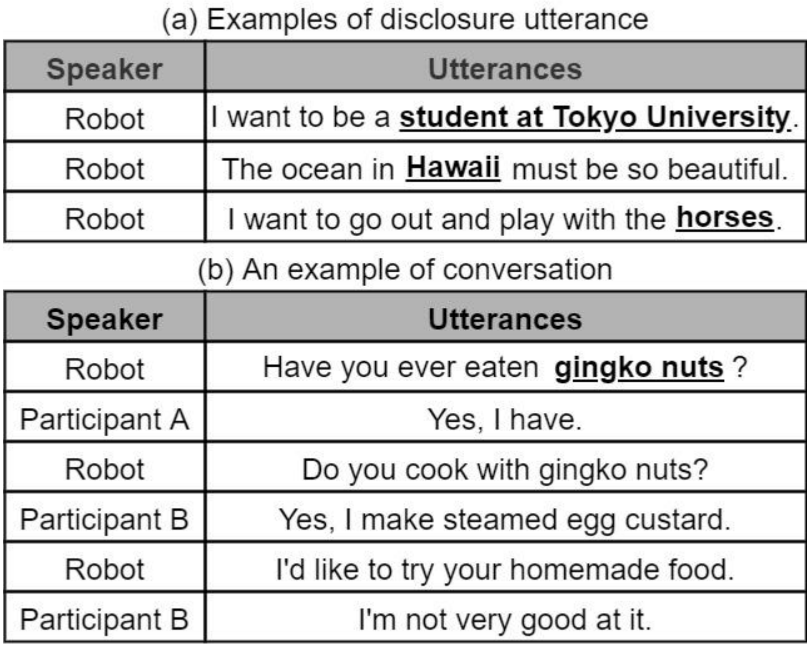
The effect of the presence of the robot on the atmosphere (Q1-2) was high, with a mean value of 5.1 and a standard deviation of 1.3. Approximately 72% of the participants felt that the presence of the robot had a relaxing effect on the TV-viewing environment.

The results of the questionnaire indicate that the robot’s utterances did not have a strong effect of causing people to look at the TV (Q1-3), with a mean score of 3.2 and a standard deviation of 1.4. From our analysis of the actual conditions of the experiment, we found that the robot's utterances and interaction tended to make the participants pay attention to the robot, and the robot's utterances did not have the effect of making them turn their attention to the TV. In fact, contrary to our intention, there were many cases wherein the robot's utterances caused the viewer to look away from the TV screen, which had a negative effect on TV viewing. This is an issue that needs to be addressed in the future.

The effect of the presence of the robot on the enjoyment of TV viewing (Q1-4) scored moderately well with a mean value of 4.5 and a standard deviation of 1.4. Approximately 56% of the respondents answered that they felt more interested in watching TV than usual due to the presence of the robot, whereas approximately 22% answered that they did not feel interested. The results of the experiment showed that some participants enjoyed the robot's utterances, whereas others felt that the robot interrupted good scenes of the TV programs. In the future, we believe that it will be necessary to adjust the timing of the robot's utterances in consideration of the importance of the TV program’s scene.

Next, we examined the participants’ impressions of the robot's disclosure utterances, conversation, and gestures (Q2). Examples of disclosure utterances and a conversation are shown in Table 2.

Regarding the robot's disclosure utterances (Q2-1), 22% of the participants found the robot "interesting," and 50% found it "sometimes interesting," indicating that together more than 70% of the participants found the robot's disclosure utterances interesting. On the other hand, there were some negative comments such as that the utterances "did not match the TV program" (47%) and that the robot "said the same thing over and over again" (44%).



(a) Examples of disclosure utterance

Speaker	Utterances
Robot	I want to be a **student at Tokyo University**.
Robot	The ocean in **Hawaii** must be so beautiful.
Robot	I want to go out and play with the **horses**.

(b) An example of conversation

Speaker	Utterances
Robot	Have you ever eaten **gingko nuts** ?
Participant A	Yes, I have.
Robot	Do you cook with gingko nuts?
Participant B	Yes, I make steamed egg custard.
Robot	I'd like to try your homemade food.
Participant B	I'm not very good at it.

Table 2 – Examples of disclosure utterances and a sample conversation. Keywords are underlined

Regarding the conversation with the robot (Q2-2), 38% of participants said they "enjoyed" the conversation and 38% said they "sometimes enjoyed" it, which indicates that more than 70% of the participants enjoyed their conversation with the robot when watching TV. In addition, 44% of the participants said that they "enjoyed watching their friends talking with the robot," even when they were not interacting with the robot themselves. On the other hand, as with the disclosure utterances, 38% of the negative comments were that the robot’s utterances "did not match the TV program" and 63% responded that the robot "said the same thing over and over again."

Lastly, regarding the robot's movements and gestures (Q2-3), 34% of participants responded that they were "interesting" and 28% responded that they were "sometimes interesting," 56% responded that they were "cute," and 22% responded that they were "sometimes cute," both of which were positive evaluations. In addition, compared to disclosure utterances and conversations, the results for the robot’s movements showed fewer negative comments.

About experimental environment

In this experiment, we tried to keep the participants in a relaxed environment, just as they would usually experience when watching TV. We investigated whether the participants were able to participate in the experiment in a relaxed and enjoyed manner (Q3). The results of the survey are shown in Figure 8.

The overall rating for enjoyment of the experiment (Q3-1) was high, with a mean score of 6.0 with a standard deviation of 1.1. The reasons given for the enjoyment were "I enjoyed watching TV with my friends and family," "I was able to re-watch TV programs I had previously watched," and "I was able to watch TV programs I wanted to watch." Therefore, we can conclude that participants were able to enjoy their time during the experiment.

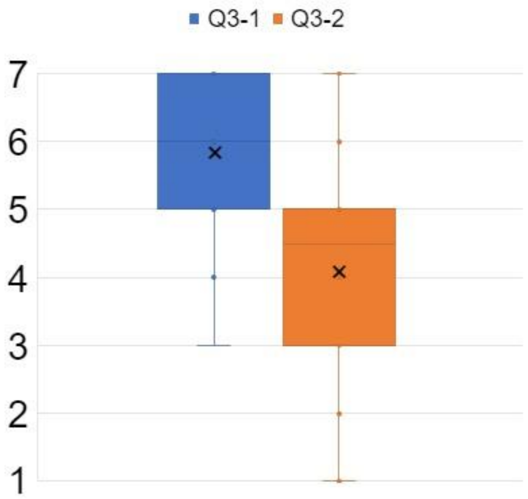


Figure 8 – Results of questionnaire on the experimental environment

Alternatively, regarding tiredness resulting from the experiment (Q3-2), the mean value was 4.1 with a standard deviation of 1.5, indicating that participating in the experiment made some people feel tired. Many of the participants cited "watching TV for a long time" as the reason for feeling tired, indicating that the length of the experiment, rather than the presence of the robot, made those participants feel tired, even though we conveyed that they could take breaks freely.

Taken together, we can conclude that although some participants felt tired due to the length of the experiment, overall, most participants enjoyed the experiment and were in a relaxed state.

CONCLUSION

In this study, we presented the functional requirements for a robot that watches TV with people, based on the analysis of the dialogue that occurs between people when watching TV. In addition, we validated the effects of a prototype robot, developed as the first step in achieving a robot based on these functional requirements.

The results of the experiment demonstrated that approximately 70% of the participants responded that "the presence of the robot increased the conversation between us," or "the presence of the robot made the atmosphere more relaxed." This indicates that the presence of the robot in the TV-viewing environment promoted communication and that users find the experience more interesting than that of their usual TV viewing. In summary, we provided evidence that the introduction of robots into the TV-viewing environment has the potential to produce novel and enjoyable media experiences.

ACKNOWLEDGMENTS

We would like to thank KDDI Research, Inc. for their cooperation in implementing the dialogue function with the robot and operating the joint experiments.

REFERENCES

1. United Nations, 2019. Patterns and trends in household size and composition: Evidence from a United Nations dataset.

2. Hoshi, Y., Kaneko, Y., Uehara, M., Hagio, Y., Murasaki, Y., Nishimura, S. and Yamamoto, M., 2020. Utterance function for companion robot for humans watching television. IEEE International Conference on Consumer Electronics. pp. 1 to 5.

3. Ogawa, H., Ikeo, M., Ohmata, H., Yamamura, C. and Fujisawa, H., 2018. System architecture for IoT services with broadcast content. IEEE International Conference on Consumer Electronics. pp. 1 to 2.

4. Goto, J., Kim, Y., Miyazaki, M., Komine, K. and Uratani, N., 2003. A spoken dialogue interface for TV operations based on data collected by using WOZ method. Annual Meeting of the Association for Computational Linguistics. pp. 101 to 104.

5. Nishimura, S., Kanbara, M. and Hagita, N., 2019. Atmosphere sharing with TV chat agents for increase of user's motivation for conversation. International Conference on Human-Computer Interaction. pp. 482 to 492.

6. Vstone Co.,Ltd., CommU. https://www.vstone.co.jp/products/commu/index.html.

7. Nakadai, K., Okuno, H. G. and Mizumoto, T., 2017. Development, deployment and applications of robot audition open source software HARK. Journal of Robotics and Mechatronics. vol. 29, No. 1, pp. 16 to 25.

8. Microsoft, Microsoft Azure. https://azure.microsoft.com/.

9. Kudo, T., 2006. MeCab: Yet another part-of-speech and morphological analyzer. http://taku910.github.io/mecab/.

10. Sato, T., Hashimoto, T. and Okumura, M., 2017. Implementation of a word segmentation dictionary called Mecab-ipadic-NEologd and study on how to use it effectively for information retrieval (in Japanese). Annual Meeting of the Association for Natural Language Processing. NLP2017–B6–1.

11. Ren, S., He, K., Girshick, R. and Sun, J., 2017. Faster R-CNN: Towards real-time object detection with region proposal networks. IEEE Transactions on Pattern Analysis and Machine Intelligence. vol. 39, No. 6, pp. 1137 to 1149.

12. Lin, T. Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P. and Zitnick, C. L., 2014. Microsoft COCO: Common objects in context. European Conference on Computer Vision. pp. 740 to 755.

13. Hou, Q., Cheng, M., Hu, X., Borji, A., Tu, Z. and Torr, P. H. S., 2019. Deeply supervised salient object detection with short connections. IEEE Transactions on Pattern Analysis and Machine Intelligence. vol. 41, No. 4, pp. 815 to 828.

14. Amazon, Amazon Web Service. https://aws.amazon.com/.

15. Mikolov, T., Sutskever, I., Chen, K., Corrado, G. and Dean, J., 2013. Distributed representations of words and phrases and their compositionality. Proceedings of the 26th International Conference on Neural Information Processing Systems. vol. 2, pp. 3111 to 3119.

16. Wu, J., Naito, M., Hoashi, K.s and Takishima, Y., 2019. A chatbot AI corresponded with TV program (in Japanese). ITE Annual Convention. 33B-2.