Adaptive Interview Strategy Based on Interviewees Speaking Willingness Recognition for Interview Robots

Fuminori Nagasawa, Shogo Okada, Takuya Ishihara, and Katsumi Nitta

Abstract—Social signal recognition techniques based on nonverbal behavioral sensing allow conversational robots to understand the user's social signals, thereby enabling them to adopt interaction strategies based on internal states inferred from the social signals. This research investigates how the online social signal recognition and adaptive dialog strategy influences the dynamic change in a user's inner state. For this purpose, we develop a semiautonomous interview robot system with an online speaker's willingness recognition module and an adaptive question selection module based on the willingness level. The online recognition model of speaker willingness is trained from multimodal nonverbal features extracted using a novel interview corpus, which allows appropriate interview questions to be chosen based on the estimated willingness level of the user. We conduct the experiment using the system to evaluate the effectiveness of adaptive question selection based on the willingness recognition model. First, the multimodal willingness recognition model is evaluated using the interview corpus. The best recognition accuracy of willingness level (high or low) was 72.8% with the random forest classifier. Second, 27 interviewees were interviewed with the two interview robot systems: (I) with the adaptive question selection module based on willingness recognition and (II) with a random question selection strategy. The proposed adaptive question strategy significantly increased the number of utterances with high willingness compared with the baseline system (II); thus, adaptive question selection with online willingness recognition elicited the speaker's willingness even though the model cannot be estimated with near-perfect accuracy.

Index Terms—Human robot interaction, Interview agent, Multimodal machine learning, Social signal processing, Speaker's willingness

1 Introduction

Recent developments in nonverbal behavior recognition techniques enable systems to recognize social signals and social behavior [1], such as turn taking, agreement, politeness, and engagement in social interaction. Many previous works have focused on analyzing the various types of social signals observed in different communication settings (monologue to audience, dyadic and small group) and multimodal nonverbal behaviors. The findings from these studies have been used to apply social signal processing (SSP) techniques in conversational agents and robots. SSP plays a central role in dialog management for conversational agents or robots in an open environment [2] and in user engagement estimation for adapting the dialog strategy [3]. One of the main challenges is to develop an adaptation mechanism for a spoken dialog system to recognize the user's inner state, such as the user's sentiment, and to adapt the dialog strategy accordingly. One ultimate goal is for the system to elicit user behavior and statements through user interaction based on adaptation techniques.

In this paper, we describe an interview robot system with social signal sensing and adaptation of the interview strategy. The core technology in this robot system is the adaptive strategy of interview questions based on the results of recognition of the interviewee's speaking willingness (inner state estimated from social signals).

Applications of interview dialog include motivational interviews, life logging, and interviews for documentary production. These are called "qualitative" or "in-depth" interviews [4], as they elicit rich and deep answers that are embedded in the personal stories told by the interviewee, rather than just answers to preprepared questions. In such applications, it is important to motivate the user to provide more information based on the interviewee's speaking willingness.

A common objective of interviews is to elicit information from interviewees by asking appropriate questions [4]. Therefore, the interviewer, who asks questions in the interview, is expected to receive emotional and social signals from the interviewee during the dialog and to motivate the interviewee to participate in the interview. One approach to motivate an interviewee is to explore a topic in depth while inviting the interviewee to spontaneously disclose information. Based on the importance of the self-disclosure of interviewees, Soleymani et al. [5] analyzed the multimodal behaviors of self-disclosing interviewees and found that the linguistic content of verbal behavior and head gestures such as nods and speech pauses were also associated with self-disclosure.

One of the most important interviewing techniques is to follow up on a topic through further questions about the topic. Following up on a topic gives the interviewee the impression that the interviewer is interested in him/her and encourages spontaneous disclosure of information. However, if the interviewee is not interested in the topic, follow-up will decrease the interviewee's willingness to participate in the interview. In such a case, the interviewer should change topics to find other topics that the interviewee is

interested in discussing [4].

Therefore, to conduct an appropriate in-depth interview that elicits the interviewee's willingness to talk, it is important to capture the interviewee's attitude and willingness to speak during the dialog. Based on theoretical findings, we developed an interview robot that adopts a topical interview strategy by asking questions based on the speaker's willingness recognition results.

First, the recognition model of user willingness is trained with multimodal audio-visual features, and the recognition model outputs the estimated willingness label per interviewee's answering utterance. Second, interview questions are adaptively chosen from a tree-structured question set based on the results of the willingness recognition model. When the interviewee answers question (i) with high willingness, a question on the same topic as (i) is chosen in the next turn. When the interviewee answers with low willingness, a question on a different topic is selected. We conduct a user study using an interview robot system with an adaptive question strategy based on the willingness recognition model (multimodal SSP model). The experimental results indicate that the adaptive strategy with willingness recognition can increase the number of utterances with high willingness. In addition, we analyze the relationship between the recognition accuracy of willingness and the number of utterances with high willingness. The main contributions are summarized as follows:

Online speaker willingness recognition in the HRI setting: We address the novel challenge of developing a prediction model of the willingness level of an interviewee. Willingness in the interview is determined by the interest level in the questions or the motivation to answer a question. We collected a multimodal corpus of human-robot interview interactions to develop a recognition model of user willingness in the interview setting. To apply the model in an interview robot system, The model is trained to recognize the willingness level per turn using audio-visual multimodal features extracted in an online manner.

Development of an interview robot system based on SSP: We develop an interview robot system with the online willingness recognition model and adaptive question selection based on the estimated willingness level. The robot system can interview users in an almost automatic manner, including online willingness recognition and adaptive question selection. Only the start time of the question is controlled by the system operator. The adaptive question selection strategy is useful to automatically conduct interviews that elicit rich and deep answers [4] embedded in the personal stories of the interviewee, such as life-logging and documentary production. The effectiveness of interviews conducted with the adaptive question selection strategy is evaluated through a user study.

Evaluation of the effectiveness of SSP in HRI: The main challenge in this paper is to analyze the impact and influence of online social signal sensing on user behavior in conversations. The interview robot system with online willingness recognition enables us to analyze the influence of social sensing. We compare the user's impression and behavior between the interview setting with the adaptive question selection strategy and the setting without the proposed strategy. Through interview interaction experiments with 27 interviewees, We show that adaptive question selection based on willingness level recognition can increase the number of utterances with high willingness, even though the multimodal willingness recognition model is not perfect (recognition accuracy is approximately 75%). The evaluation process of the social signal sensing module on HRI can be applied to other applications.

The rest of this paper is organized as follows. Section 2 describes related work. Section 4 presents the interview robot system with the speaking willingness recognition model. Section 5 discusses the multimodal interview corpus used to train the willingness recognition models. Section 6 describes how the speaker willingness recognition model is trained based on multimodal features. The experimental setting for evaluating the system is described in Section 7, and The results are presented in Section 8. Finally, the results are discussed in Section 9, and The research is concluded in Section 10.

2 Related works

This research is related to a multimodal dialog system, and the core technique is multimodal social signal processing for utilizing a communication robot or embodied conversational agent (empathetic agents).

2.1 Engagement and willingness

In human-agent or robot interactions, many studies [2], [3], [6] focus on engagement recognition based on users' multimodal behaviors. Engagement is defined as an attitude that determines the quality of interaction in [7]. The main difference between "willingness" in this research and "engagement" is that willingness denotes an inner state of whether the participant would like (desire) to talk about the interviewer's questions and does not denote an attitude such as engagement. The attitude observed from interviewees with a high willingness level is sometimes similar to the attitude of those with a high engagement level, so we review research analyzing engagement to clarify the similarities and differences between willingness and engagement.

Engagement also represents how much a user is interested in and willing to continue the current dialog [6]. Nakano et al. [3] proposed a method for estimating whether the user is engaged in the conversation based on gaze transition patterns of the user's gaze sensing behavior. Gaze behavior patterns when the user was distracted from the conversation were also analyzed. Inoue et al. [6] proposed a recognition model of user engagement in human-robot interactions using a hierarchical Bayesian model that estimates both the user's engagement level and the annotator's character as latent variables. The character represents a template for the perception of engagement correspondence. For example, annotators with one character tend to regard laughing behavior as the engagement indicator.

In the real world, where multiple interviewees can come and go, the system must estimate who is interacting with the system or when the user is interacting with the system.

Bohus et al. [2] proposed a multiparty engagement recognition model for predicting engagement based on visual analysis. They developed open-world conversational

systems that operate in relatively unconstrained environments where multiple participants might come and go, establish, maintain and break the communication frame, and simultaneously interact with a system and with others. In their system, the robot senses the position of the person coming from various directions and the robot's position and uses them as features to estimate engagement. In contrast, we developed an interview robot system that operates in an environment where the robot and interviewee interact 1-on-1 and no third party intervenes in the conversation. To this end, we focused more on a single interaction subject, measured the person's postural features as more detailed features than the positional relationship between the robot and the person and used acoustic features extracted from speech utterances for willingness estimation.

Sidner et al. [8] defined the concept of engagement as "the process by which interactors start, maintain, and their perceived connections to each other during an interaction". Bohus et al. [2] and Nakano et al. [3] used the definition in [8] to annotate the engagement level in their research. Oertel et al. [9] classified the definitions of engagement used in related works and concluded that engagement is the attitude observed as a result of interest in dialog, sustained attention, concentration, and participation. As they point out, engagement has been used to refer to a number of related but different concepts.

A common definition of engagement is a person's active attitude toward his or her interaction partner or his or her statements when the person is a speaker or listener. In this study, we wanted to examine willingness to disclose information (i.e., providing additional information) in interviews, but engagement has multiple definitions and is too broad in meaning. Therefore, we constructed and annotated a willingness scale based on findings from previous interview studies.

2.2 Social signal processing for HAI/HRI

We introduce related research on social signal processing, mainly its application to human agent/robot interaction (HAI/HRI). Hirano et al. [10] presented a multimodal modeling method with multitask learning to recognize multiple labels, such as interest levels, sentiment levels and next-action decisions, to implement adaptation strategies for multimodal dialog systems. Hirano et al. [10] enhanced the multitask learning framework utilizing weakly supervised learning (WSL) algorithms for which the target label is not necessarily accurate.

Virtual agents with social signal sensing have recently been developed for communication skills training; Mohammed et al. developed the dialog system "MACH" for training job interviews. They conducted a one-week interview training for students using MACH, and the students' interview performance was evaluated by human experts. The results showed that students who interacted with MACH were rated as having improved overall interview performance [11]. Tanaka et al. developed a dialog system that teaches social communication skills through dialog with people with autism spectrum disorder (ASD) [12] and, for automatic training of social skills, the user's listening skills during a conversation with a computer agent. They proposed an assessment of user listening skills during conversations with computer agents for automated social skills training [13] and developed a computer avatar with spoken dialog to observe the communication behavior of participants with dementia [14]. Several studies have focused on the detection of user interests and concerns. Hirayama et al. developed a concurrence system based on eye gaze and speech analysis in which the system provides detailed information and recommendations according to the user's interests [15].

Chiba et al. estimated the user's level of interest in the dialog content from the user's nonverbal behaviors, such as the acoustic spectrum of speech, positional characteristics of each facial part, and eye movements during speech. Araki et al. created a corpus of dialog data for the study of dialog and user interest. [16] Tomomasu et al. proposed a method to determine whether a user is interested in a particular topic using facial expression recognition and prosodic information of speech utterances [17].

Batrinca et al. analyzed Big-five personality trait recognition in human-robot interaction settings. The results showed that cooperative behavior caused subjects to develop traits related to sociability (e.g., agreeableness and extraversion), and uncooperative behavior caused them to develop traits related to anxiety (e.g., emotional stability/neuroticism) [18].

Weber et al. [19] developed a dynamic user modeling approach based on reinforcement learning that enables a robot to analyze a person's reaction while the robot tells jokes and continuously adapts its sense of humor. Nasihati et al. [20] presented dialog management routines for a system to engage in multiparty agent-infant interactions. The system measures attention by means of an eye tracker and measures patterns of emotional arousal using a thermal infrared imaging camera. A dialog policy is presented to select individual actions and plan multiparty sequences based on perceptual inputs about the infant's internal changing states of emotional engagement. Saito et al. [21] developed a turn-taking mechanism based on recognizing the subject's attitude toward speaking up or not speaking up as an agent to interview elderly people with dementia.

DeVault et al. [22] presented a virtual human interviewer system designed to create engaging face-to-face interactions in which the user feels comfortable talking and sharing information. The key technique is adapting the agent's nonverbal behavior based on recognition of the multimodal behavior of users, including facial expressions and acoustic features [23]. In particular, the system in [22] was designed to create interactional situations that are favorable to the automatic assessment of distress indicators, defined as verbal and nonverbal behaviors correlated with depression, anxiety or posttraumatic stress disorder (PTSD). Simsensei predicts the next action based on verbal and nonverbal information of the user. In contrast, our system uses only nonverbal behavior. Soleymani et al. [5] analyzed verbal and nonverbal behavior during intimate self-disclosure. They trained a multimodal deep neural network to estimate the level of self-disclosure. Correlation analysis of verbal and nonverbal behavior revealed that the linguistic content of verbal behavior is associated with self-disclosure. Overall, word count, verbally expressed affective and cognitive pro-

cesses and sentence construction were important indicators of intimate self-disclosure. Head gestures such as nods and speech pauses were also associated with self-disclosure.

Kobori et al. [24] developed a text-based interview dialog system and showed that the system's ability to engage in small talk unrelated to the interview questions enhanced the user's impression of the dialog. Our research focuses on the changes in the interviewee's willingness that occur as a result of choosing the content of the dialog itself. Chiba et al. [25] investigated a method for estimating the interviewee's willingness to continue the dialog from multimodal features, with the goal of making the interview dialog last longer. In the study, willingness was defined as "the desire for speaking continuity" or "the desire to disclose the information one has". They analyzed interview dialog conducted by human interviewers, but we consider the change in interviewee's willingness when using a robot as an interviewer. Ishihara et al. [26] proposed a recognition model of the interviewee's willingness in the interview interaction based on multimodal behavior (i.e., verbal, audio, and visual). To establish an interview robot that can adapt the interview strategy by recognizing an interviewee's willingness, we develop and evaluate a real-time willingness recognition model and an adaptive interview strategy based on estimated willingness.

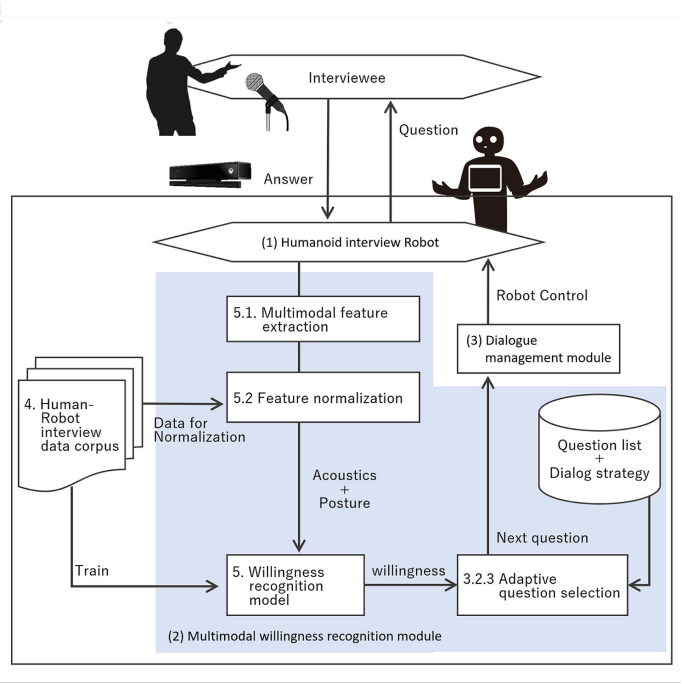


Fig. 1. Interview robot system

3 Difference from related works

The main difference between our research and previous research proposing a robot or agent with social signal recognition models is summarized as follows. First, we develop an interview robot with an adaptive question selection strategy based on speaking willingness-level (social signal) recognition and evaluate the strategy. Multimodal modeling for online speaking willingness recognition in the human-robot interview setting has not been well explored, and investigating the effectiveness of adaptive question selection based on willingness recognition is a first challenge. Although Inoue et al. [27] proposed a method to generate follow-up questions based on the spotting of proper nouns as the focal point in user utterances, they did not focus on adaptation based on social signal sensing. Second, we evaluate the effectiveness of the proposed adaptive strategy based on SSP via a user study including both the amount of behavioral change of users (an objective evaluation) and a questionnaire survey (a subjective evaluation). Some previous research, such as [19], has shown that social signal sensing and adaptation (optimization) of a robot's behavior based on the sensing result improves the user's experience of dialog with the robot (system) through questionnaire surveys. We focus on evaluating not only the impression of users toward the dialog experience with the system but also how the online social signal sensing per utterance affects the user's inner state or attitude dynamically. Finally, we show that adaptive question selection based on the estimated willingness level alters user behavior (acoustic and visual activity) and elicits utterances with high willingness levels.

4 Interview robot system based on SSP

An overview of the proposed interview robot system with a social signal (speaker's willingness level) recognition module is shown in Fig. 1.

This section describes the humanoid conversation robot (Section 4.0.1), the sensing environment for the interview robot system (Section 4.1) and the interview interaction scenario and adaptive question selection based on the willingness recognition results (Section 4.2).

The proposed interview robot aims to elicit information from the interviewee through an adaptive question selection strategy. Fig. 2 shows the configuration of the interview robot and interview dialog system. The proposed interview robot is composed of the following: (1) the humanoid interview robot, (2) the multimodal willingness recognition module, and (3) The dialog management module.

4.0.1 Humanoid conversational robot

The interview robot is composed of a human-shaped personal robot and a multimodal sensing system. The personal robot, named Pepper [1], was developed by SoftBank Mobile Corp and has speech synthesis and smooth hand and head motion generation modules. Pepper is 1.2 m tall and weighs approximately 30 kg.

Pepper is associated with module (1) as an interviewer to interact with the interviewee. Willingness recognition and question selection are performed by module (2) on the backend. The backend module (2) consists of a multimodal sensing module, a willingness recognition module, and a question selection module. The speech synthesis and gestures in Pepper are automatically controlled by NAOqi SDK [28]. Module (3) is responsible for sending the question selected by module (2) to Pepper by calling the text-to-speech function of Pepper SDK.

Thus, the multimodal sensing module, willingness recognition module, and question selection module control

1. https://www.softbank.jp/robot/consumer/products/

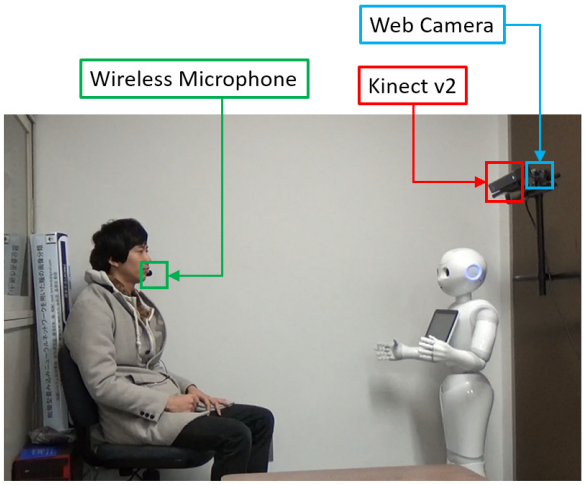


Fig. 2. Interview scene with the interview robot system

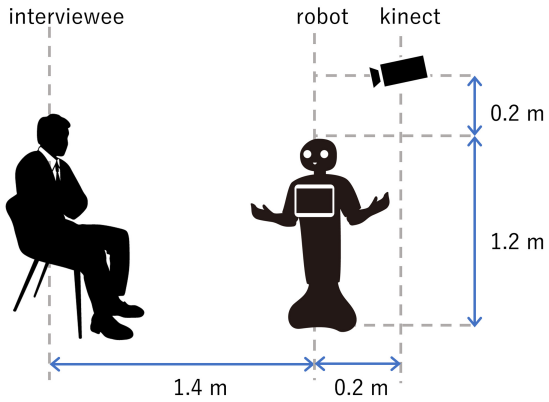


Fig. 3. Layout of the interviewee, the robot, and the Kinect sensor

the humanoid interview robot module. The multimodal behavior sensing module, willingness recognition module, and question selection module of interviewee were implemented for this study.

The multimodal sensing module is used to estimate willingness from the multimodal data observed while the interviewee is answering. The system selects the next question and transfers it to Pepper.

4.1 Multimodal sensing environment

We collected the interviewee's multimodal data using a web camera (logicool C910, 1080p 30 fps), Kinect V2 and wearable microphone (Shure PGA31 headset microphone) during the interview dialog (Fig.2). The arrangement of the participants, the robot, and the sensors is shown in Fig. 3. The interviewee sits 1.4 m in front of the robot, and the webcam and Kinect sensors are placed $0.2m$ above the robot's head and $0.2m$ behind the robot. The interviewee and these sensors face each other across a distance of $1.5m$. We train the recognition model of the interviewee's willingness from the coordinates of the joints estimated by the Kinect sensor and the audio collected by the wearable microphone. Audio and visual features are computed, and the computed features are used to learn a recognition model of the interviewee's willingness. The multimodal features of speech and vision and the recognition model of willingness are explained in Section 6.1.

4.2 Adaptive question selection based on the interviewee's willingness level

We propose a question selection module based on the recognized willingness level. The question list is composed of a hierarchical tree structure, as shown in Fig. 4. Each node denotes one question in the interview. The next question is selected by moving to another node from the current node on the structure.

Each node is linked to two nodes: (i) a node on one lower layer and (ii) a node on the same layer. (i) A node on one lower layer denotes a more detailed question on the same topic as the current one. (ii) A node on the same layer denotes a question on a different topic. If the system decides to switch the topic of the question, node (ii) is referred to as the next "current question", and the system asks the question of node (ii) as the next one.

Whether the next node (question) is (i) or (ii) is based on the willingness recognition result shown in Fig. 4. A red circle denotes a recognition result of "high willingness" and a black cross denotes a recognition result of "low willingness" If the recognition result for the previous interviewee utterance (answer to the previous question) is characterized by "high willingness" the system asks a question (i) to follow up on the topic, and if it is characterized by "low willingness" the system asks a question (ii) to change topics. The details of the dialog strategy based on multimodal willingness recognition are described as follows.

Tree search methods The questions on the tree-structured list are selected by switching the two search methods (depth-first search and breadth-first search) [29] for the tree structure.

Depth-first search Depth-first search gives priority to the children of the current node. If the current node has child nodes, the child nodes are selected. If it does not have any child nodes, it moves to the parent node and performs the same search. This process is performed recursively to select the next question.

Breadth-first search The breadth-first search prioritizes nodes in a shallow hierarchy. A sibling node of the current node (a subnode of the parent node other than the current node) is selected. If no sibling node is found, a sibling node of the parent node is chosen. This process is repeated recursively to select the next question.

Using these two search methods, the developed system selects the next question in the following steps:

- Step 1 Multimodal data are recorded while the interviewee is answering a question.
- Step 2 The multimodal features extracted from the data (recorded in Step 1) are input into a trained model for willingness recognition.
- Step 3 The willingness level is determined based on the multimodal features.
- Step 4 Step 4-a or Step 4-b is performed according to the output from the willingness recognition result.

Step 4-a (If the utterance is recognized as high willingness) Select the next question by a depth-first search starting from the current question. Specifically, question (i) one layer below the current question is selected as the next question. If there is no lower node for the current question, perform Step 4-b.

Step 4-b (If the utterance is recognized as low willingness) The next question is selected by a breadth-first search starting from the node one higher than the current question. Specifically, the lower node (i) of the current question is invalidated, and question (ii) in the same hierarchy as the current question is selected.

Step 5 Ask the selected question and return to Step 1.

With this question selection flow, we can conduct interviews with any question scenario, as long as we have a list of questions with a similar tree structure.

TABLE 1
Questions used for the experiment

No.	contents
1.	What is your current research theme?
2.	When did you start the current research?
3.	Why did you choose the current research theme?
4.	What was the most enjoyable event in your research?
5.	What are you having trouble with in conducting research?
6.	What is the appeal of your current research?
7.	How is it applied to your research?
8.	What are you interested in besides research?
9.	What was your previous research theme?
10.	What was the result of the previous research?
11.	Which is more fun between the current and past research?
12.	Why do you think so?
13.	What are your hobbies?
14.	What do you care about in balancing private life and research?
15.	Please tell me your impression of this dialogue.

5 HUMAN ROBOT INTERVIEW DATA CORPUS

We collected a human-robot interview corpus to train the willingness recognition model of the interview robot system. The corpus is collected as training data for the willingness recognition model of the interview robot.

5.1 Corpus setting

To collect this data corpus, we recruited 8 interviewees (7 male/1 female, aged 22-30 years). The Research Ethics Committee of the Tokyo Institute of Technology reviewed and approved the collection of data and the corresponding research using this dataset. The interview robot system asked questions in order based on the prepared list shown in Table 1. The eight interviewees were graduate school students, so the interview topic was "research topic majored in graduate school".

The start time of each utterance of the robot was decided by an operator. During the interview session, multimodal data, including audio speech data and depth image data, were recorded. The multimodal data were automatically segmented per exchange, which consisted of a system utterance (question) followed by an interviewee utterance (answer to the question) using the start and end times of the system utterance. The eight interviewees were each interviewed once, so a total of eight sessions were collected.

5.2 Willingness level annotation

The willingness in the interview was determined by the interest level regarding the questions or speaking motivation caused by their interest level to the question. The willingness label is annotated per interviewee's turn. The system needed to estimate the willingness level per turn to make the decision of whether to change the current topic of the question. The system's turn, the interviewee's turn, and the willingness annotation interval are shown in Fig. 5. A willingness-level label is annotated per turn, so the total number of exchanges (the paired question from the robot and answer from the interviewee) corresponds to the number of samples. We defined utterances in turn with high willingness as those in which the interviewee was interested in the question and had an attitude of providing additional information.

Low-willingness utterances were defined as simplified answers or answers that avoided explanations of specific content.

We asked three coders to watch the videos of the interviews and annotate the interviewees' willingness or unwillingness when answering the questions. Coders were instructed to consider various features of the participants, such as body activity, acoustic and utterance content, and not to determine the labels only with a specific modality.

We provided the annotators with instructions for examples of high/low willingness. In the case of "high willingness", the interviewee not only answered the question but also provided additional answers, such as a detailed explanation of the related field, his/her own experiences, or a personal theory. In contrast, in the case of "low willingness," the interviewee seemed to cut off their answers after a short response or avoid explaining specific details.

First, these coders annotated the willingness level using a 5-point Likert scale (lowest willingness: 1 to highest willingness: 5). Second, the average values $\overline{v}$ of levels $\{v_1, v_2, v_3\}$ annotated by three coders were converted into binary values by using threshold point 3 (corresponding to neutral). This means that samples with an average value greater than 3 ($\overline{v} > 3$) were categorized into the high-willingness class, and those with a value smaller than 3 ($\overline{v} \leq 3$) were categorized into the low-willingness class.

In this study, the particularly highly motivated sample was classified as a high-willingness class, while the rest of the sample was classified as a low-willingness class. Thus, 3 (neutral) was classified in the low-willingness class. Willingness is an inner state that is not completely observable from external information, so We need to analyze how difficult it is to annotate the score by human coders. We calculated the agreement for the original willingness score (1 to 5) between the annotators using the weighted kappa. The weighted kappa was $\kappa_w = 0.91$, indicating sufficient agreement.

It might be difficult to correctly annotate willingness as "the desire for dialog continuity" in a general interaction setting (e.g., casual chatting) because the roles (speaker or listener) of the interlocutors change dynamically and

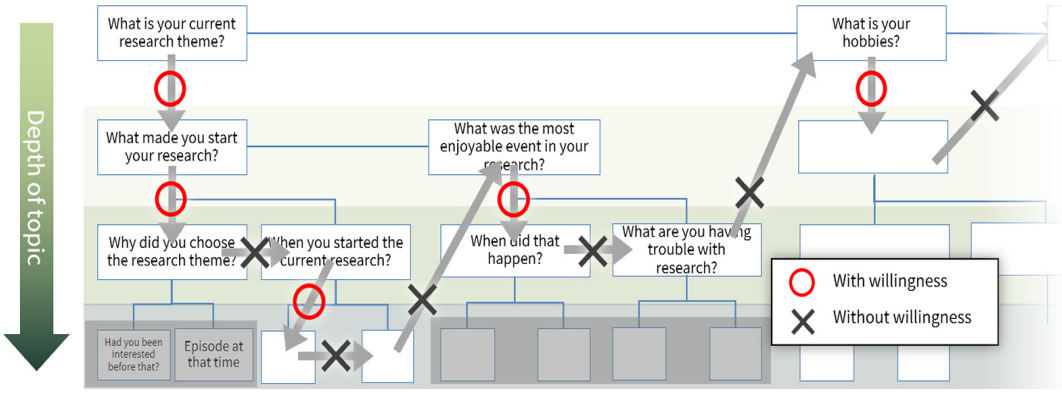


Fig. 4. Example of adaptive choice based on estimated willingness. Each question node, represented by a box, is arranged in a tree structure. Based on the estimated willingness, the next question is selected from this tree structure.

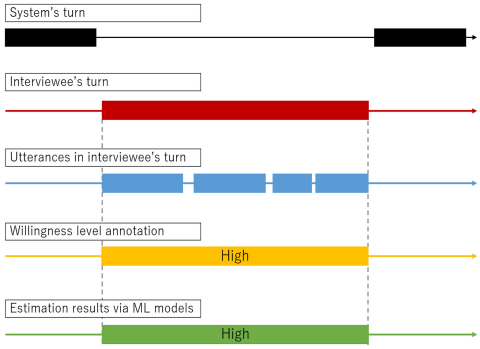


Fig. 5. The section that performs the robot's turn, the interviewee's turn and the willingness annotation

the observed multimodal features are varied in such a conversation setting. Conversely, the role of the speaker is constrained as an interviewee in the interview setting in this study. Annotators can compare the willingness levels of speakers among QA turns. As the annotation result is affected by the constraint in the interview task, we consider that high agreement is obtained.

6 Willingness Recognition Model

The willingness recognition result is used to select the next question, so the model was trained to infer the willingness level per exchange in an online manner. The input data to the model are composed of the multimodal behavioral features that are observed while the user is speaking to answer the current question. The model outputs the willingness level (high/low) corresponding to the input multimodal features.

To determine whether the system changes the current topic in the next question, we set the willingness recognition problem as a binary classification task of willingness level (high or low). The binary willingness recognition model is trained with the annotated willingness label and the multimodal behavioral features observed while the user is speaking.

6.1 Multimodal feature extraction

Acoustic features were extracted from the speech signal obtained from a microphone. Posture features were extracted from the three-dimensional coordinates of each joint of the upper body, which were estimated via Microsoft Kinect v2. The total number of dimensions of the features was 139.

6.1.1 *Acoustic features*

Acoustic features in speech signals and speaking status represent the inner state of a speaker, such as emotion. First, we extracted the speech length of each answer utterance as the speech timing feature under the hypothesis that if the user speaks with high willingness, they will answer the question with a longer speaking time.

We used OpenSMILE [30] to calculate the acoustic features. The acoustic features include the root mean square frame energy (RMSenergy), mel frequency cepstral coefficient 1-13 (MFCC 1-13), and fundamental frequency (F0). In addition, we used speaking length. Finally, 4 statistics, namely, the mean, standard deviation, minimum and maximum, were calculated and used as acoustic features. The total number of acoustic features was 61 $(15 \times 4 + 1(speakinglength))$.

6.1.2 *Posture features*

We used the three-dimensional coordinates of each joint of the upper body, estimated via Microsoft Kinect v2, to extract posture features. In this study, we used posture data of the head, shoulders, elbows, hands, thumbs, and hand tips of both the right and left arms.

We calculated 2 statistics, mean and variance, of the time series of coordinates observed while the user was speaking, as well as acoustic features. The total number of posture features was 78.

6.1.3 *Feature analysis with Student's t test*

We investigated the features that contribute to classifying the topic continuance labels based on a statistical t test. The objective of a t test is to test the hypothesis that the means of samples in the binary classes of each feature are equal. The acoustic and posture features that were significantly

TABLE 2
Features with significant differences between low-willingness and high-willingness by t-test

p-value	Features (Posture)	Features (Audio)
5%	Shoulder Y Position (mean) Shoulder Z Position (mean) Elbow Y Position (mean) Shoulder norm (mean)	speech length pitch (max) energy (min) MFCC (mean)
2%		pitch (min) pitch (mean)

different ($p < 0.05$) between the high/low willingness groups are listed in Table 2.

Acoustic Features: Six acoustic features, namely, speech length, minimum energy, mean MFCC, minimum pitch and maximum pitch, were significantly different. Regarding the acoustic features, all features that were significantly different showed higher values in "high-willingness" situations.

Posture Features: Four posture features were significantly different: the mean values of Shoulder_Y, Shoulder Z, Elbow_Y, and Shoulder_norm. There was also a significant difference in the mean value of Shoulder_norm, which is the distance between the shoulder coordinates and the measurement origin. The mean value of each coordinate in the high-willingness case is smaller than that in the low-willingness case. On the other hand, for the variance, no characteristic was significantly different between the high-willingness group and the low-willingness group. Since these position values are expressed as the distance from the Kinect sensor, this result indicates that when willingness is high, the interviewee's posture tends to be closer to the sensor, that is, the interviewee leans forward.

For the elbow and shoulder postural features, significant differences were found for either the right or left values, but whether this was left or right varied between interviewees. Additionally, no significant difference was observed for the left and right values added together. The reason for this may be that the interviewee's posture tends to change or the interviewee tends to answer by moving his or her hands (body) when willingness is high.

6.2 Feature normalization for the online recognition task

Normalizing features to reduce the influence of individual differences, such as the physique and acoustic characteristics of the interviewee, is important for improving the social signal recognition accuracy from multimodal features. In this study, the nonverbal features were normalized to the range of $[0.0, 1.0]$ using a min-max normalization method. Let $x(t, d)$ be the value of the dth dimension in the multimodal feature vector corresponding to the tth exchange. The minimum value is $X_{min}(d)$, and the maximum value is $X_{max}(d)$ for all features observed from an interviewee in an interview session. Thus, the normalized feature value $x_n(d)$ is obtained according to the following equation:

$$x_n(t, d) = \frac{x(t, d) - X_{min}(d)}{X_{max}(d) - X_{min}(d)} \tag{1}$$

The min-max normalization method can be used only for training data because it requires all exchanges in a session; the method cannot be used for test data because the willingness level is estimated per exchange in an online manner and all exchanges cannot be used for normalization.

To address this problem, we propose an approximate normalization method to normalize the test data. This method assumes that the range of values for each feature in the training data is approximately similar to the range of values in the test data. First, for the training data, each feature is normalized within samples observed from an interviewee using the equation 1.

In the training phase, the calculated range of the feature value ($X_{max} - X_{min}$) per interviewee is stored, and the average range is used to normalize the test data. Let $x(k, t, d)$ be the value of the dth dimension in the feature vector corresponding to the tth exchange of interviewee k in the training dataset. The minimum value $X_{min}(k, d)$ and maximum value $X_{max}(k, d)$ represent the values over all exchanges. The range $r(k, d)$ of the value of the dth dimension of interviewee k is calculated as $r(k, d) = X_{max}(k, d) - X_{min}(k, d)$. $x(k', t', d)$ of the test data, which is the value of the dth dimension in the t'th exchange of unknown interviewee k', is normalized to x_n using the following equation:

$$x_n(k', t', d) = \frac{x'(k', t', d) - X_{min}(k', d)}{\bar{r}}, \quad \bar{r} = \frac{1}{N_t} \sum_k r(k, d) \tag{2}$$

In this equation, N_t is the number of training samples.

6.3 Machine learning model

In this study, interviewees willingness was estimated from multimodal data using machine learning. We trained two machine learning models, random forest and support vector machine (SVM), and the accuracy of each model was evaluated via cross-validation. The model with the best estimation accuracy was used for the adaptive interview dialog system.

6.3.1 Linear support vector machine (SVM)

In the binary classification task, linear SVM models [31] based on acoustic, posture and multimodal features were trained to compare the estimation accuracy. We used the SVM in early fusion (EF) to fuse the different modalities. In EF, the feature vectors from different modalities were concatenated into one feature vector. In the SVM model, the final estimation was based on the decision function of the unimodal models.

6.3.2 Random forest

As a comparative method, we used random forest in EF to fuse the different modalities, similar to the aforementioned SVM modeling.

7 Experimental Settings

First, we evaluated the binary classification models of the willingness labels trained with machine learning models and the external annotation score (average of scores by

annotators). The objective of the first experiment was to validate how accurately the willingness level can be predicted using the multimodal features.

Second, we evaluated the interview robot system with the online willingness model. through interview interaction sessions between the robot and interviewees. The objective of the second experiment was to evaluate the effect of the adaptive strategy on the willingness level of the interviewees.

7.1 Evaluation of the willingness recognition model

To validate the accuracy of willingness recognition, we trained the SVM model and random forest model and evaluated the trained models as follows.

Training models: The SVM models were optimized using a cross-validation scheme for the training dataset with the penalty parameter set as $\{0.001, 0.01, 0.1, 1, 10\}$. The penalty parameter ensures a balance between the loss function and margin maximization. In the random forest model, the number of trees was set to $\{1, 10, 100, 1000\}$, and there were no restrictions on the maximum number of leaf nodes or the maximum tree depth. The model was optimized using a cross-validation method on the training data.

Evaluating models: Leave-one-person-out cross-validation (LOPOCV) was used to evaluate the trained models for willingness recognition. In LOPOCV, the test data corresponded to the samples observed in the interview sessions of one interviewee, and the remaining samples from the other interviewee were used as training data. We report the average accuracy of the test dataset (Section 6.2).

7.2 Evaluation of the adaptive interview strategy

The first objective of this experiment is to evaluate the effectiveness of the adaptive interview (question selection) strategy based on willingness recognition with the models trained in Section 7.1. The second objective is to investigate how the proposed interview strategy differentiates the willingness level of interviewees after adaptation and how it influences impressions of the interview. We conducted two interview sessions per interviewee: (I) a session with question selection by means of the adaptive strategy and (II) a session with random question selection. For each session, we compared the percentage of utterances with high willingness, which were annotated by the interviewees to validate the effectiveness of the adaptive strategy.

7.2.1 Participants

We recruited 30 participants as interviewees through a human-resource agency in Japan. Participants in the experiment were recruited from a wide range of ordinary Japanese. The participants had a 50-50 male/female ratio, and their ages ranged from 20 to 60 (mean age=39.3), with each age group evenly distributed. The participants were paid a flat fee through a staffing agency as a reward for their participation in the experiment. Before each experiment, we explained to the participant that he or she could discontinue participation in the experiment at will and that the video and other recorded data would not be released to the outside and obtained consent. During the experiment, participants were not subjected to unreasonable physical or mental strain, and the recorded video and other datasets were managed to prevent information leakage. The Research Ethics Committee of the Tokyo Institute of Technology reviewed and approved this experiment and the corresponding study using the dataset obtained in the experiment.

7.2.2 Experimental design and procedure

To evaluate the adaptive interview strategy, we asked the interviewees to be interviewed by two systems: system (I) and system (II). The only difference between the systems was the selection of the next question. System (I) conducted interviews by selecting the next question based on the proposed adaptive strategy with the willingness recognition model. System (II) conducted interviews by selecting the next question based on a random selection strategy. We call the strategy of system (I) the "adaptive strategy" and that of system (II) the "random strategy".

In the random strategy, the same binary tree structure used for the adaptive strategy is used; the system randomly decides whether to switch topics for the next question. To make it easier for interviewees to talk with the system, we generated the question list based on their favorite topics via a slot filling method.

The base question list is shown in 3. The slot " (topic)" in each question is filled with the topic selected by the interviewee before the interview. The interviewees could select the favorite topic from six topics: sports, hobbies, study, research, work, and childcare. The experiment was conducted according to a within-subjects design. All subjects participated in the experiment under both conditions. The order in which the interviewees were interviewed with systems (I) and (II) was randomly decided to prevent an effect of order on the interviewees' behavior.

7.2.3 Measures

We evaluated the effectiveness of the proposed adaptive strategy based on SSP via a user study including both the amount of behavioral change of users (an objective evaluation) and a questionnaire survey (a subjective evaluation).

Comparison of utterances with willingness:

To investigate the effect of adaptive question selection-based willingness recognition, we compared the number of QA exchanges (a paired question and its answer) with high willingness between system (I) using an adaptive strategy and system (II) using a random strategy.

As mentioned in Section 7.2.2, each interviewee was interviewed by systems (I) and (II) once each. After each interview session, we asked the interviewees to watch a video of the interview for the two sessions and annotate their willingness levels (high or low) corresponding to the answer to each question. We directly compared the percentage of exchanges with high willingness in the entire dialog between the two strategies (adaptive vs random).

Questionnaire survey for impression of the system: We analyzed the interviewees' impressions of our system by means of a questionnaire survey. After the interview sessions, the interviewees answered the five questions listed below.

TABLE 3
Question scenario used for the experiment

No.	Depth of topic	Content
1	0	What kind of (topic) are you doing now or in the past?
2	0	What became a cause of you beginning (topic)?
2-1	1	When were the events that triggered you to start (topic)?
2-1-1	2	Could you tell me about a detailed episode?
2-2	1	What did you think about (topic) when you began?
2-2-1	2	What do you think about (topic) now compared to when you began?
3	0	Are there memories that you enjoyed about (topic)?
4	0	On the other hand, do you have any bad or painful memories related to (topic)?
4-1	1	How did you overcome an issue when it occurred?
5	0	Is there anyone you met through (topic)?
5-1	1	Please tell me about the episode that got you acquainted with that person.
6	0	What do you like about the (topic).
6-1	1	(About the answer to question 6) Why do you like it?
6-1-1	2	(About the answer to Question 6) When do you realize what you like about it?
7	0	Conversely, what kind of things do you dislike about (topic)?
7-1	1	(About the answer to question 7) Why do you dislike it?
7-1-1	2	(About the answer to Question 7) Do you sometimes feel bad about disliking that characteristic?
8	0	What kind of things are you conscious of in the future to continue (topic)?
8-1	1	(About the answer to question 8) For that, what do you want to do specifically?
9	0	Is there anything else you would like as a new challenge in the field of (topic)?
9-1	1	(About the answer to Question 9) When did you know that?
9-1-1	2	(On answer to question 9) How did you know about it?
9-2	1	(About the answer to question 9) Why did you decide to try that challenge?
9-2-1	2	(About the answer to question 9)When are you planning to challenge?
9-3	1	(About the answer to Question 9) Are you making concrete plans etc. for actually challenging?
9-3-1	2	(On answer to Question 9-3) Have you talked to someone about the plan to challenge?
9-4	1	(About the answer to Question 9) Do you know anyone already doing that challenging field?
10	0	Finally, what is (topic) in your life? or what does (topic) mean in your life?

CQ1 Did you feel that the robot was interested in your answers in the interview? (attitude of interest)
CQ2 Did you feel that the robot was asking questions about topics that you are happy to answer? (unpleasant question)
GQ1 Did you feel it was easy to talk with the robot compared to talking to people? (ease of talking)
GQ2 Did you feel anything was strange about the dialog?
GQ3 (If you felt strange) What was the degree of discomfort? (degree of discomfort)

The questions consisted of two comparison questions (CQ) and three general questions (GQ). CQ1 and CQ2 were used to quantitatively evaluate the dialog strategies and were asked once for each dialog strategy. The answers to CQ1 and CQ2 are explained in Section 8.2.2. GQ1, GQ2, and GQ3 were used to clarify the limitations and future work of the system and were asked once throughout the entire dialog; the answers to GQ1, GQ2, and GQ3 are explained in Section 9.

Questions CQ1, CQ2, and GQ1 were rated on a 5-point scale (1: agree, 2: slightly agree, 3: undecided, 4: slightly disagree, 5: disagree). Question GQ2 was a binary-choice question (1: yes, 2: no), and question GQ3 was a five-point evaluation of the intensity of discomfort (1: very much bothered, 2: somewhat bothered, 3: undecided, 4: somewhat not bothered, 5: hardly bothered). We also asked the interviewees who answered "1: I am concerned" in GQ2 to write down the specific aspects that made them feel uncomfortable.

7.2.4 Analysis

The objective of the analysis was to clarify the effectiveness of the adaptation strategy with willingness recognition. **Testing hypotheses for validating the adaptive strategy:** We investigated two hypotheses on the effectiveness of the proposed adaptation strategy. The first hypothesis is that interviewees will continue to speak with high willingness if the system accurately recognizes their willingness level and continues to ask relevant questions. To validate this hypothesis, we compared the percentage of utterances with high willingness during the dialog session for each of the two strategies (adaptive vs. random). The results are described in Section 8.2.1. Our second hypothesis is that if the system accurately recognizes willingness levels and continues to ask questions in a way that keeps high-willingness topics and changes the current interviewing topics based on the detection of low-willingness QA, it can improve the interviewee's impression of the interview dialog. To investigate the interviewees' impressions of the interview session, We asked participants whether the robot was interested in the interviewee's answer (CQ1) and whether the robot asked an unpleasant question (CQ2). To test this hypothesis, we compared the distribution of respondents for both questions (CQ1,2) between the two strategies (adaptive vs. random) by using a statistical t test to determine whether there was a significant difference. The results are described in section8.2.

Case studies: We analyzed the relationship among willingness recognition accuracy, the impression score of the questionnaire, and willingness level in representative interview sessions as case studies. We analyzed the case of interviewees whose percentage of willingness was lower when the adaptive strategy was used.

TABLE 4
Results of the cross-validation test. The test was performed for each combination of acoustic and posture features using two classifiers.

	classifier	Acoustic (A)	Posture (P)	A+P
All features	SVM	69. 9	46. 3	72. 8
	Random Forest	66. 9	45. 6	71. 3
Selected features	SVM	61. 8	61. 8	62. 5
	Random Forest	61. 0	60. 3	61. 8

TABLE 5
Results of cross-validation of each normalization method. The highest accuracy is achieved by "full-normalized". "Approximate-normalized" and random forest are more accurate than "non-normalized".

Normalization	SVM	Random Forest
Full-normalized	72. 8	71. 3
Approximate-normalized	53. 6	68. 6
Non-normalized	52. 2	50. 7

TABLE 6
Comparison of the number of utterances and the percentage of utterances with high willingness for different dialogue strategies

	Percentage of utterances with high willingness[%]	Number of utterances
Random strategy	43. 1	17. 26
Adaptive strategy	55. 5	13. 52
T-test result	0. 002	0. 005

8 RESULTS

8.1 Accuracy of willingness estimation

We compared the accuracy of models trained in various conditions (unimodal and multimodal features, machine learning methods) to find the optimal model to recognize the willingness level. Table4 shows the classification accuracy of the willingness estimation models.

8.1.1 Comparison between multimodal features:

In terms of the comparison between the unimodal models (acoustic or posture), Columns 3 and 4 in Table4 show the accuracy of the unimodal model with acoustic (A) and posture features (P). The best accuracy of 69.9% was achieved by the SVM model with acoustic features. The random forest model with acoustic features also obtained better accuracy (66.9%) than the model with posture features. According to these results, acoustic features are effective in classifying the willingness level, regardless of the machine learning model.

Column 5 of Table4 shows the accuracy of the multimodal model (A+P). Both SVM and random forest with multimodal features (A+P) obtained better accuracy (72.8%, 71.3%) than the best unimodal models. The results show that fusing acoustic and visual features improved the recognition accuracy.

8.1.2 Effect of approximate normalization:

As noted in Section 6.2, Our robot system requires an online recognition model to select the next question based on the recognition result of the willingness label. For the online recognition model, we present the normalization method working on the condition that the ranges of feature values are unknown for normalizing the multimodal features observed from an unknown (new) interviewee. In this section, We analyze the influence of the approximated normalization method on the recognition accuracy.

We compare the approximated normalization method with a complete normalization method (fully normalized) using the range of feature values of the test data and a method without normalizing both the training and test data (nonnormalized). In realistic situations, the range of the test data from a new interviewee is unknown, so we cannot use the fully normalized method for the online recognition task in the robot system.

Table 5 compares the recognition accuracy. The best accuracy is obtained by the fully normalized approach (71.3% in random forest, 72.8% in SVM). Although the accuracy of the approximated method was degraded with respect to that of the fully normalized approach, The approximated method obtained an accuracy of 68.6% in random forest. The decrease in accuracy was limited to 3.8%. The accuracy is 17.9% better than that of the nonnormalization method. The results show that the approximated method can mitigate the degradation in accuracy by means of the difference in the range of the test data. Finally, the best accuracy in the online recognition setting was obtained by the random forest model with the multimodal feature set, so the multimodal random forest classifier with approximated normalization was utilized in the interview robot system.

8.2 Evaluation of the proposed strategy's efficiency

In this section, we present the results obtained from the experiments described in Section 7.2, which are based on quantitative measures.

8.2.1 Comparison of utterances with high willingness

Table 6 shows the number of utterances and the percentage of utterances with high willingness. Column 2 of Table 6 shows the percentage of utterances with high willingness. The percentage of utterances with high willingness was higher when the adaptive strategy (55.5%) was used than when the random strategy (43.1%) was used. Conversely, the percentage of exchanges shown in column 3 of Table 6 indicates that the number of utterances was lower for the adaptive strategy than for the random strategy. We conducted t tests to evaluate the significance of the difference in the "percentage of utterances with high willingness". We obtained $p < 0.05$ for both the "Percentage of utterances with high willingness" and "Average number of exchanges" results.

The percentage of willingness of each interviewee is shown in Fig. 6. In the case of the adaptive strategy, the percentage of willingness was higher for 21 of 27 individuals. Fig. 6 shows that the 21 interviewees tended to speak with high willingness more often when the adaptive strategy was used.

8.2.2 Questionnaire survey for impression of the system

Table 7 and Fig. 7 show the results of the questionnaire conducted in Section 7.2.

Rows 3 through 7 show the number of people who chose each option for each question, and row 8 shows the weighted average of the number of people who responded for each strategy by option number. row 9 shows the 95%

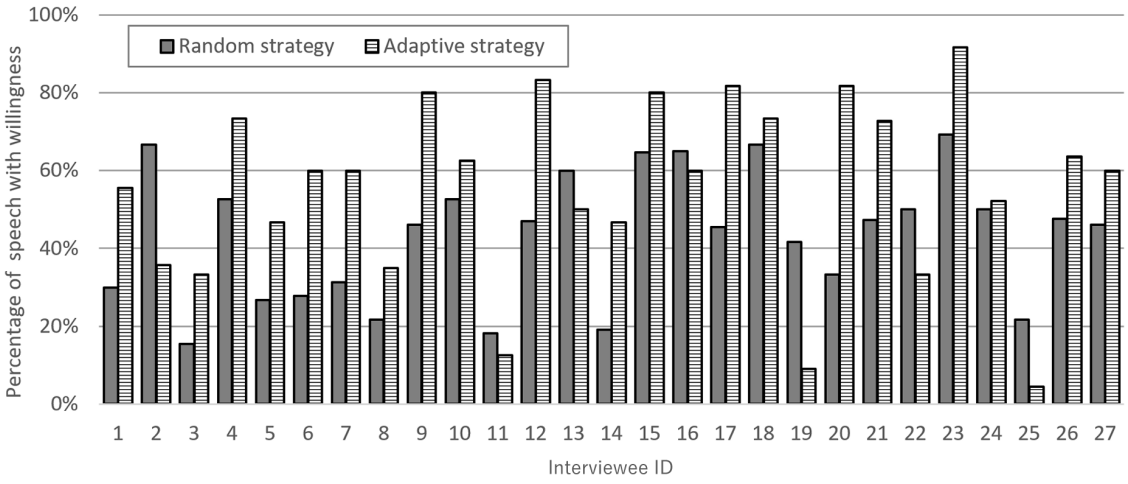


Fig. 6. Percentage of "with high willingness" utterances per interviewee.

TABLE 7
Results for CA1 (answers to CQ1) and CA2 (answers to CQ2) of the questionnaire in the interview experiment (unit: persons)
CQ1:"Did you feel that the robot was interested in your answers in the interview ? (attitude of interest)"
CQ2: "Did you feel that the robot was asking questions about topics that you were happy to answer? (unpleasant question)"

	CA1 (small is better)		CA2 (large is better)	
	Adaptive	Random	Adaptive	Random
1: Aagree	6	4	0	0
2: Slightly agree	13	11	5	7
3: Undecided	1	3	8	8
4: Slightly disagree	6	9	8	10
5: Disagree	1	0	6	2
Mean	2. 37	2. 63	3. 56	3. 26
95% interval	0. 47	0. 44	0. 42	0. 37
Effect size	$d = 0.23$		$d = 0.30$	
T-test result	0. 025		0. 067	

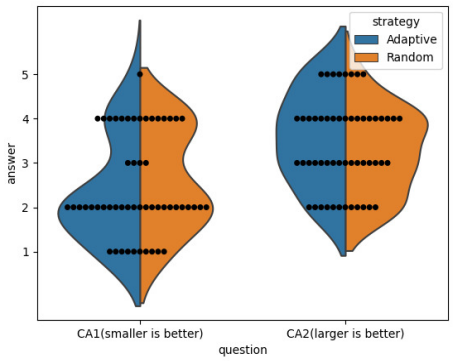


Fig. 7. Violin plots showing the distribution of respondents for CA1 (responses to CQ1) and CA2 (responses to CQ2) of the questionnaire in the interview experiment. The number of respondents for each option is represented by black dots.

confidence interval, row 10 shows the effect size for each question between the adaptive strategy and random strategy, and row 11 shows the t test result for each question between the adaptive strategy and random strategy. Columns 2 and 3 show CA1, the answer to CQ1; since CA1 is a question about the strength of positive impressions, 1 (agree) is the best answer, and 5 (disagree) is the worst answer. Columns 4 and 5 show CA2, the response to CQ2; since CA2 is a question about the strength of negative impressions, 1 (agree) is the worst impression, and 5 (disagree) is the best impression.

The averages of the questionnaire ratings show that CA1 was rated higher in the adaptive strategy and CA2 was rated higher in the random strategy.

The distribution in Figure 4 shows that for CQ1, the distribution on the side of smaller values is larger for the adaptive strategy than for the random strategy; for CQ2, the distribution on the side of 5 is smaller for the random strategy than for the adaptive strategy.

In the t test results, there was a significant difference in CA1. This result shows that adaptive question selection based on estimated willingness allows the system to give the impression of being more interested in the interviewee's speech.

We compared the depth of the interviews in the two strategies: we compared the depth of reaching the maximum depth for each topic for the questions in Table 3 for which there was at least one question with a depth of topic of 1 or more. The results showed that the average was 0.48 for the random strategy and 0.53 for the adaptive strategy, but the t test result was p=0.18, which was not significantly different. This does not mean that significantly deeper topical questions were asked in either of the two dialog strategies. Nevertheless, the results in Table 6 show a higher value for the "percentage of willingness" and the results in Tables 7 through 9 show that the interviewees' impressions of the dialog improved as a result of appropriate topic selection by the adaptive strategy.

8.2.3 *Impressions of adaptive interview dialog*

Table 8 shows the number of people who responded to each option and the weighted average by number for question GQ1. The answer with the largest number of respondents was 3 ("undecided"), indicating that for the majority of interviewees, the robot did not give the impression that it was extremely easy or difficult to talk to compared to humans. Table 9 also shows the number of respondents for each option for the question about whether they felt any

TABLE 8
The number of respondents for each option to the question GQ1 (larger is better). The most common answer was "undecided", indicating that the robot did not give the impression of being extremely easy or difficult to talk to compared to a human.

	Num. of people
1: Agree	0
2: Slightly agree	4
3: Undecided	12
4: Slightly disagree	6
5: Disagree	5
Mean	3. 44
95% confidence interval	0. 39

TABLE 9
The number of people who responded to the question about discomfort with the dialogue in the survey. The largest number of respondents chose "somewhat not bothered", indicating that most interviewees were not bothered by uncomfortable content in the dialogue.

	Num. of people
0: (There was no discomfort)	6
1: Very much bothered	3
2: Somewhat bothered	3
3: Undecided	1
4: Somewhat not bothered	11
5: Hardly bothered	3
Mean	2. 63
95% confidence interval	0. 73

strangeness in the dialog or the intensity of the strangeness. In Table 9, the number of respondents who answered "no" to question GQ2 is assigned to option "0 (There was no discomfort)". Table 9 shows that the most common answer was "4 (somewhat not bothered)", indicating that many interviewees did not feel much discomfort with the dialog content.

In GQ3, we asked the respondents who answered that they felt uncomfortable in GQ2 to describe the specific points that they felt uncomfortable with. Topic clustering was performed on the responses obtained from the interviewees in free-text form. As a result, four topics common to several interviewees were extracted. Representative examples of responses belonging to the four extracted topics are listed in GA3-a through GA3-d.

- GA3-a It was difficult to grasp the meaning of some questions, or the questions were unnatural.
- GA3-b The system repeated the same question.
- GA3-c When I felt that the next question I answered was not truly relevant, I felt that the robot was not listening to me.
- GA3-d It was a long time between the answer and the next question.

GA3-a suggests that the quality of the questions for the keywords prepared by the system was insufficient. In this paper, the questions were created by applying the topics to the predesigned question templates shown in Table 3. This result shows the limitation of question generation via the template. Keeping the topic alive through the automatic generation of questions may be useful for solving this problem. GA3-b and GA3-c show the necessity of using speech recognition and natural language processing for question selection. GA3-b was provided by interviewees who talked ahead of what they were going to be asked in the next question, and GA3-c was provided by an interviewee who experienced switched topics by the system when the end of the in-depth question tree was reached. GA3-d shows the challenges of processing speed for willingness estimation and question selection. In the willingness estimation process of the system presented in this paper, the calculation of multimodal features took at least 1 second. In addition to overcoming the challenges of natural language processing described above, accelerating the process of willingness estimation is also an important future work.

8.2.4 Case study

We analyzed the case of interviewees whose percentage of willingness was lower in the case of the adaptive strategy. Fig. 8 shows the estimated willingness level and the ground-truth label annotated by the interviewee (low or high). In addition, the recognition accuracy for the willingness estimation and the content of the responses to the questionnaire are described. In each graph, the horizontal axis denotes the elapsed time in the dialog, and the willingness level (high or low) is plotted on the vertical axis. The left side of the figure shows the percentage of each interviewee's motivation and their responses to the questionnaire (CA1 and CA2).

Four cases are shown in Fig. 8. ID 22 and ID 16 are examples where the percentage of willingness is lower for the adaptive strategy. ID 9 and ID 23 are examples with a higher percentage of willingness exchanges in the case of the adaptive strategy and are shown for comparison. Accuracy was low for ID 22 and ID 16 and high for ID 9 and ID 23. If the accuracy was low in all four cases, factors other than accuracy likely changed the intention rate and responses to the questions, but the results of the present study showed that the two cases with high accuracy and the two cases with low accuracy showed different trends for the percentage of willingness and answers to the questionnaire. These results suggest that the higher the accuracy of the willingness estimation, the higher the percentage of utterances with high willingness.

The graphs in the timeline showed a discrepancy between the true value and the estimated value (i.e., false negative error) immediately before the true value changed from high to low. This trend was common to all interviewees, which suggests that it is difficult to identify a change from high to low. This topic will be addressed in future research.

On the comparison of results of the questionnaire, most of the errors (false negatives) in ID 22 and ID 23 estimated the utterances with high willingness as low willingness. On the other hand, for ID 16 and ID 9, who had no false-positive errors, the results of CA1 were higher than those of ID 22 and ID 23. This suggests that the false-positive error in the willingness estimation worsened the CA1 scores. On the other hand, ID 22 and ID 23 showed not only false-negative but also false-positive errors (errors in estimating high willingness for low-willingness utterances) compared to ID 16 and ID 9. Although ID 23 had higher accuracy and percentage of willingness, their CQ2 score in the questionnaire survey was worse than that of ID 9. This suggests that

false-positive errors in the willingness estimation worsen the CQ2 score.

9 Discussion

We discuss the limitations of the proposed adaptive strategy with willingness level recognition and the robot system to clarify the remaining work.

9.1 Effectiveness of the adaptive strategy

Tables 6 and 7 show that the proposed adaptive question selection strategy based on willingness recognition achieves better results in the subjective evaluation of users than that of random question selection. The results show the effectiveness of adaptive question selection, which continues asking questions on topics that the user has high willingness about and stops asking questions on topics that the user has low willingness about.

In interview interactions, It is important for the interviewer to elicit more information and self-disclosure from the interviewee. Kobori et al. [24] analyzed the effect of ice-breaking dialog (unrelated to interviews) in interview interactions on the text dialog system and found that ice-breaking dialog influences users's impressions. Chiba et al. [25] presented the recognition model of willingness to talk using the interview interaction data corpus collected by the Wizard-Of-Oz (WoZ) method to analyze the factors for continuing the dialog while maintaining the user's desire for dialog continuity.

Compared to these related studies, the novel findings are that adaptive question selection improved users impressions of the interview experience and significantly increased the number of utterances with high willingness levels. On the engineering side, a contribution of this research is the development of a semiautonomous interview robot [2] equipped with the multimodal willingness estimation model and adaptive question selection. With the interview robot, we could conduct experiments to investigate the adaptive question strategy based on willingness recognition.

A future direction for developing the adaptive strategy is to identify a mechanism for eliciting more various kinds of information from users through interview interaction. Hiramaya et al. [15] proposed a proactive interaction strategy called "mind probing" to elicit user reactions. The central idea in human-system interaction is to sense the reaction behaviors of users to the system's act after a prior act from the system side to estimate the user's internal state. They introduced a digital signage system as a prototype system. First, the system highlights a region (corresponding to the system's prior act) on the signage display. Second, the system estimates the user's interest level in the highlighted region based on sensing the eye gaze activity (reaction behavior) of the user to the region. The study [15] shows that the highlighting act by the system elicits the user's reaction and makes the automatic estimation of interest level accurate. This proactive strategy is a reference for our future work. It is important to investigate the appropriate design of the question strategy or nonverbal behavior of robots to elicit user reactions or answers to improve the user's willingness estimation performance.

2. The start time when the robot asks questions is controlled by an operator.

9.2 Significance of the adaptive interview robot

The advantage of the adaptive interview robot is supported by the findings of [4]. Ben et al. [4] discussed the advantages and disadvantages of interviews by comparison with questionnaire surveys. Among the advantages, when more than a couple of open questions are asked, an interview is less burdensome as the respondent's workload. Conversely, the questionnaire is quite a burden for respondents because they are forced to do a lot of writing to answer the questions adequately. Among the disadvantages, an interview does not permit anonymity due to the simple fact that an interviewer is present. In addition to the anonymity issue, the interviewee often adapts an answer so that it conforms to the interviewer's values and preferences. The proposed interview system is useful to mitigate the disadvantages of interviews because The system does not have an interview strategy based on specific values and preferences and selects appropriate questions based on the willingness level of the interviewee. The implicit motivation of the system design is to elicit what they would like to talk about with the interviewer. It is also very important to avoid continuing to ask questions that interviewees do not feel like answering (with low willingness).

The aim of most interviews is to obtain answers to the questions that are relevant to the interviewer's goal. Willingness estimation is not essential in all interviews. However, willingness estimation is important in interviews for life-logging and interviews for documentaries. In such interviews, the key role of the interviewer is to listen to the interviewee and to elicit what the interviewee would like to talk about by encouraging their self-disclosure. Mohammad et al. [5] developed a deep learning algorithm to automatically estimate the level of intimate self-disclosure from verbal and nonverbal behavior in interviews using human-agent interaction datasets. The question set used in the interview setting in this study is related to self-disclosure because these questions are related to the interviewee's own experience. We find that adaptive question selection based on willingness estimation increases the number of answers with high willingness to questions that promote self-disclosure.

9.3 Limitations and future work

In this research, we defined utterances with high willingness as a state in which the interviewee is interested in the question and has a positive attitude toward responding to the question. The goal of this project was to elicit more information by asking questions to follow up on the topics that the user was interested in discussing.

9.3.1 Accuracy of the willingness recognition model

As shown in Table 5, the willingness recognition model has an accuracy of 68.6% in the binary classification task. Although this estimation accuracy is higher than chance, the model fails to estimate nearly 30% of the instances. However, the results in Table 6 indicate that following up on topics based on our model increases the percentage

Interviewee ID	Acc.(%)	Percentage of Willingness(%)	Answer to Questionnaire CA1 (small is better)	CA2 (large is better)
22	36%	33%	4	3
16	50%	63%	1	4
9	60%	80%	1	5
23	66%	92%	2	4

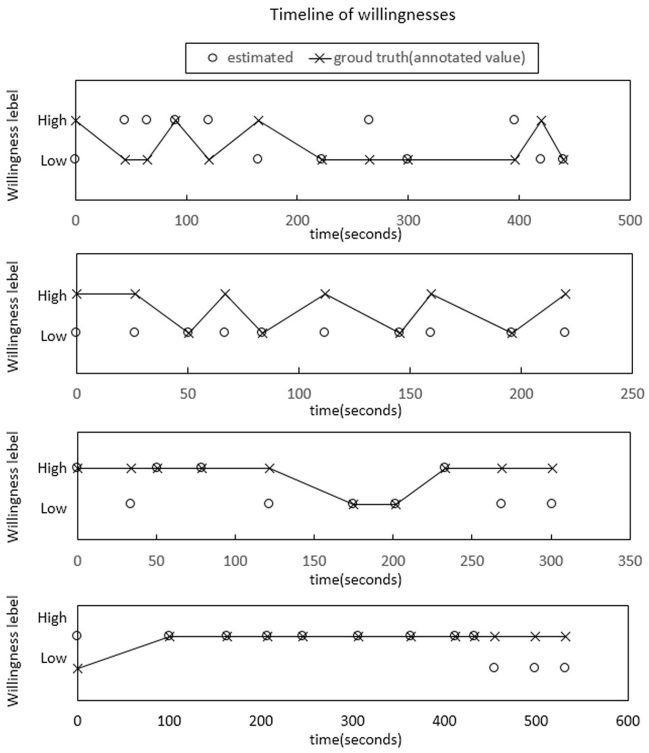


Fig. 8. The right side of the figure shows the time trends of the estimated and ground-truth willingness; the left side shows the percentage of willingness and responses to the questionnaire (CA1 and CA2). Two examples are shown for each interviewee with high/low percentages of utterances with high willingness: the low group is ID22 and ID16, and the high group is ID9 and ID23.

of utterances with high willingness and has a significant impact on the evaluation by the questionnaire survey. These results suggest that the current accuracy is effective for determining whether to follow up on a topic. By increasing the accuracy of the estimation, we expect to further increase the percentage of utterances with high willingness.

In this study, willingness estimation was performed using only basic features that are compatible with online processing. To improve the accuracy of the estimation, future work will add more detailed acoustic and facial features within the range of processing speeds that allow online recognition to improve the accuracy.

In this study, we used binary classification to estimate willingness for the purpose of controlling topic continuation/switching. We believe that estimating willingness at multiple levels using a regression model would allow for more sophisticated question selection. This is a subject for future work.

9.3.2 Follow-up on topics based on willingness estimation

In this study, questions arranged in a tree structure were prepared in advance as dialog scenarios. Therefore, it was not possible to develop and explore the topics flexibly according to the topics and answers selected by the interviewees.

As Table 6 shows, the number of response utterances was lower in the adaptive strategy than in the random strategy. Because the question scenario we prepared for this experiment had at most three layers in the question tree, even when the system followed up on a topic where high willingness was obtained, it quickly and easily reached the questions at the end of the tree. Therefore, even if topics with high willingness are followed up, the questions will be completed soon, the topic will be changed, and the question will be cut for topics with low willingness. This is the reason why the questions were completed earlier when the adaptive strategy was used than when the random strategy was used.

Inoue et al. [27] proposed a mechanism for generating in-depth questions based on analyzing words contained in the questions via automatic speech recognition (ASR) and spoken language processing (SLP). Generating adaptive follow-up questions based on ASR and SLP is a future task.

10 Conclusion

This research investigated how the adaptive dialog strategy based on online social signal recognition influences the dynamic change in the interviewee's inner state. For this purpose, we developed a semiautonomous interview robot system with an online speaker's willingness recognition module and adaptive question selection module based on the willingness level. The robot system can conduct interviews in an almost automatic manner with online willingness recognition and adaptive question selection.

First, we evaluated the multimodal willingness recognition model using the interview corpus. The online recognition accuracy for the willingness level (high or low) was highest, 68.6%, when using the random forest classifier. Second, 27 interviewees were interviewed with the two interview robot systems: (I) with the adaptive question selection module based on willingness recognition and (II) with a random question selection strategy. The proposed

adaptive question strategy significantly increased the number of utterances with high willingness. These results show that adaptive question selection with online willingness recognition elicited the speaker's willingness even though the model cannot be estimated with near-perfect accuracy. A future step toward realizing interview agents that can elicit more information from users is to combine the adaptive question selection strategy based on social signal processing and adaptive question generation based on automatic speech recognition (ASR) and spoken language processing (SLP).

REFERENCES

[1] A. Vinciarelli, M. Pantic, and H. Bourlard, "Social signal processing: Survey of an emerging domain," *Image and vision computing*, vol. 27, no. 12, pp. 1743–1759, 2009.

[2] D. Bohus and E. Horvitz, "Learning to predict engagement with a spoken dialog system in open-world settings," in *Proc. Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL)*. Association for Computational Linguistics, 2009, pp. 244–252.

[3] Y. I. Nakano and R. Ishii, "Estimating user's engagement from eye-gaze behaviors in human-agent conversations," in *Proc. International Conference on Intelligent User Interfaces (IUI)*. ACM, 2010, pp. 139–148.

[4] B. Emans, *Interviewing: Theory, techniques and training*. Routledge, 2016.

[5] M. Soleymani, K. Stefanov, S. H. Kang, J. Ondras, and J. Gratch, "Multimodal analysis and estimation of intimate self-disclosure," *ICMI 2019 - Proceedings of the 2019 International Conference on Multimodal Interaction*, pp. 59–68, 2019.

[6] K. Inoue, D. Lala, K. Takanashi, and T. Kawahara, "Latent character model for engagement recognition based on multimodal behaviors," in *Proc. International Workshop on Spoken Dialogue Systems (IWSDS)*, 2018.

[7] N. Glas and C. Pelachaud, "Definitions of engagement in human-agent interaction," in *Proc. International Workshop on Engagment in Human Computer Interaction (ENHANCE)*, 2015, pp. 944–949.

[8] C. Sidner, C. Kidd, C. Lee, and N. Lesh, "Where to look: A study of human-robot engagement," in *Proc. International Conference on Intelligent User Interfaces (IUI)*, 01 2004, pp. 78–84.

[9] C. Oertel, G. Castellano, M. Chetouani, J. Nasir, M. Obaid, C. Pelachaud, and C. Peters, "Engagement in human-agent interaction: An overview," *Frontiers in Robotics and AI*, vol. 7, 8 2020.

[10] Y. Hirano, S. Okada, H. Nishimoto, and K. Komatani, "Multitask prediction of exchange-level annotations for multimodal dialogue systems," in *Proc. International Conference on Multimodal Interaction (ICMI)*, 2019, p. 8594.

[11] M. E. Hoque, M. Courgeon, J.-C. Martin, B. Mutlu, and R. W. Picard, "Mach: My automated conversation coach," in *Proc. International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp)*. ACM, 2013, pp. 697–706.

[12] H. Tanaka, H. Negoro, H. Iwasaka, and S. Nakamura, "Embodied conversational agents for multimodal automated social skills training in people with autism spectrum disorders," *PloS one*, vol. 12, no. 8, p. e0182151, 2017.

[13] H. Tanaka, H. Negoro, H. Iwasaka, and S. Nakamura, "Listening skills assessment through computer agents," in *Proc. International Conference on Multimodal Interaction (ICMI)*. ACM, 2018, pp. 492–496.

[14] H. Tanaka, H. Adachi, N. Ukita, M. Ikeda, H. Kazui, T. Kudo, and S. Nakamura, "Detecting dementia through interactive computer avatars," *IEEE journal of translational engineering in health and medicine*, vol. 5, pp. 1–11, 2017.

[15] T. Hirayama, Y. Sumi, T. Kawahara, and T. Matsuyama, "Info-concierge: Proactive multi-modal interaction through mind probing," in *The Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC 2011)*, 2011.

[16] M. Araki, S. Tomimasu, M. Nakano, K. Komatani, S. Okada, S. Fujie, and H. Sugiyama, "Collection of multimodal dialog data and analysis of the result of annotation of users' interest level," in *Proc. International Conference on Language Resources and Evaluation (LREC)*. European Language Resources Association (ELRA), 2018.

[17] S. Tomimasu and M. Araki, "Assessment of users' interests in multimodal dialog based on exchange unit," in *Proc. International Workshop on Multimodal Analyses Enabling Artificial Agents in Human-Machine Interaction*. ACM, 2016, pp. 33–37.

[18] L. Batrinca, N. Mana, B. Lepri, N. Sebe, and F. Pianesi, "Multimodal personality recognition in collaborative goal-oriented tasks," *IEEE Transactions on Multimedia*, vol. 18, no. 4, pp. 659–673, 2016.

[19] K. Weber, H. Ritschel, I. Aslan, F. Lingenfelser, and E. André, "How to shape the humor of a robot - social behavior adaptation based on reinforcement learning," in *Proc. International Conference on Multimodal Interaction (ICMI)*. ACM, 2018, pp. 154–162.

[20] S. Nasihati Gilani, D. Traum, A. Merla, E. Hee, Z. Walker, B. Manini, G. Gallagher, and L.-A. Petitto, "Multimodal dialogue management for multiparty interaction with infants," in *Proc. International Conference on Multimodal Interaction (ICMI)*. ACM, 2018, pp. 5–13.

[21] N. Saito, S. Okada, K. Nitta, Y. I. Nakano, and Y. Hayashi, "Estimating user's attitude in multimodal conversational system for elderly people with dementi," *AAAI Spring Symposium - Technical Report*, vol. SS-15-07, pp. 100–103, 2015.

[22] D. DeVault, R. Artstein, G. Benn, T. Dey, E. Fast, A. Gainer, K. Georgila, J. Gratch, A. Hartholt, M. Lhommet, G. Lucas, S. Marsella, F. Morbini, A. Nazarian, S. Scherer, G. Stratou, A. Suri, D. Traum, R. Wood, Y. Xu, A. Rizzo, and L.-P. Morency, "Simsensei kiosk: A virtual human interviewer for healthcare decision support," in *Proc. International Conference on Autonomous Agents and Multi-agent Systems (AAMAS)*. International Foundation for Autonomous Agents and Multiagent Systems, 2014, pp. 1061–1068.

[23] G. Stratou and L.-P. Morency, "Multisense—context-aware nonverbal behavior analysis framework: A psychological distress use case," *IEEE Transactions on Affective Computing*, vol. 8, no. 2, pp. 190–203, 2017.

[24] T. Kobori, M. Nakano, and T. Nakamura, "Small Talk Improves User Impressions of Interview Dialogue Systems," *In Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL-16)*, no. September, pp. 370–380, 2016.

[25] Y. Chiba, T. Nose, and A. Ito, "Analysis of efficient multimodal features for estimating user's willingness to talk: Comparison of human-machine and human-human dialog," *Proceedings - 9th Asia-Pacific Signal and Information Processing Association Annual Summit and Conference, APSIPA ASC 2017*, vol. 2018-February, no. December, pp. 428–431, 2018.

[26] T. Ishihara, F. Nagasawa, K. Nitta, and S. Okada, "Estimating interviewee's willingness in multimodal human robot interview interaction," *Proceedings of the 20th International Conference on Multimodal Interaction, ICMI 2018*, 2018.

[27] K. Inoue, K. Hara, D. Lala, K. Yamamoto, S. Nakamura, K. Takanashi, and T. Kawahara, "Job Interviewer Android with Elaborate Follow-up Question Generation," *ICMI 2020 - Proceedings of the 2020 International Conference on Multimodal Interaction*, pp. 324–332, 2020.

[28] S. Robotics, "Softbank robotics documentation naoqi sdk," "http://doc.aldebaran.com/2-5/index_dev_guide.html".

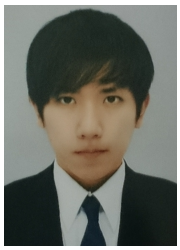
[29] D. C. Kozen, "Depth-first and breadth-first search," in *The design and analysis of algorithms*. Springer, 1992, pp. 19–24.

[30] F. Eyben, M. Wöllmer, and B. Schuller, "Opensmile: The munich versatile and fast open-source audio feature extractor," in *Proceedings of the 18th ACM International Conference on Multimedia*, ser. MM '10. New York, NY, USA: Association for Computing Machinery, 2010, p. 14591462. [Online]. Available: https://doi.org/10.1145/1873951.1874246

[31] C. Cortes and V. Vapnik, "Support-vector networks," in *Machine Learning*, 1995, pp. 273–297.

Fuminori Nagasawa received a B. Eng. degree from Aoyama Gakuin University (Shibuya-ku, Tokyo, Japan) in 2016 and an M. S. degree from Tokyo Institute of Technology (Meguro, Tokyo, Japan) in 2018. He joined Mitsubishi Electric Corporation, Information Technology Research Institute (Tokyo, Japan), in 2018 and has been a Ph.D. student at the Japan Advanced Institute of Science and Technology (JAIST) since 2020. His research interests include social signal processing, human-robot interaction, machine learning, and computer vision. He is a member of the Japanese Society for Artificial Intelligence.

Shogo Okada directs the Social Signal and Interaction Group at the Japan Advanced Institute of Science and Technology (JAIST) in Japan and is an associate professor at JAIST. He obtained his Ph.D. in 2008 from Tokyo Institute of Technology in Japan. In 2008 and 2011, he joined Kyoto University and Tokyo Institute of Technology, respectively, as an assistant professor. He joined the IDIAP Research Institute in Switzerland as a visiting faculty member in 2014. His research interests include social signal processing, human dynamics, multimodal interaction, and machine learning. He is a member of IEEE and ACM.



Takuya Ishihara was born in Shimane Prefecture, Japan, in 1993. He received a B. Eng. degree from Utsunomiya University, Utsunomiya city, Tochigi Prefecture, Japan, in 2016, and an M. Eng. degree from Tokyo Institute of Technology, Meguro City, Tokyo Metropolis, Japan, in 2018. Subsequently, he joined NTT DATA Corporation, Tokyo Metropolis, Japan, in 2018, where he worked on the research and development of object detection and tracking from drone cameras using deep learning and the self-location and shape of object estimation technology using SLAM/SfM. At present, he works as an IT consultant/engineer at Accenture, Tokyo Metropolis, Japan.

Katsumi Nitta received his B. S., M. S., and Dr. of Engineering degrees from the Tokyo Institute of Technology in 1975, 1977, and 1980. He worked as a researcher at the Electrotechnical laboratory from 1980 to 1988. During that time, he was seconded to the Institute of New Generation Computer Technology from 1988 to 1994, where he worked as a senior researcher. From 1996 to 2018, he worked for the Tokyo Institute of Technology as a professor. He is currently a specially appointed professor at Tokyo Institute of Technology, where he works on university-wide education in data science and artificial intelligence. His research interests include 'AI and law' and argumentative computer systems. He is a member of the Information Processing Society of Japan and the Japanese Society for Artificial Intelligence.