Title	内面状態の推定と対話戦略適応に基づく知的対話処理に関する研究
Author(s)	長澤, 史記
Citation	
Issue Date	2025-03
Type	Thesis or Dissertation
Text version	ETD
URL	http://hdl.handle.net/10119/19928
Rights	
Description	Supervisor: 岡田 将吾, 先端科学技術研究科, 博士

Doctoral Dissertation

Research on intelligent dialogue processing based on inner state estimation and adaptive dialogue strategies

Fuminori Nagasawa

Supervisor: Shogo Okada

Graduate School of Advanced Science and Technology
Japan Advanced Institute of Science and Technology
(Information science)

March 2025

Abstract

The ultimate goal of this research is to realize a computer system that can connect users to appropriate functions and services by empathizing with their feelings and interests through adaptive interactions and eliciting their inner needs. As a method for eliciting information from people, listening dialogue systems that use dialogues such as interviews and interviews have been actively studied. In listening dialogue systems, it is important to realize empathic dialogue, in which the system recognizes the inner state of the user and motivates the user to participate in the dialogue.

To realize such empathic dialogue processing, it is necessary to realize a mechanism in which the system proposes or changes topics according to the user's willingness, which is the user's inner state of "wanting to talk".

First, a machine learning model was developed to estimate the interviewee's willingness (desire to disclose information they have) based on their posture during the dialogue and the prosodic features of their speech utterances. Machine learning models using Random Forest and LinerSVM were trained. A method was developed to reduce the impact of individual differences in multivariate features on estimation accuracy for first-time interviewees who do not have the information necessary for normalizing multivariate features. Using the interview corpus collected through the dialogue experiment, this study evaluated the accuracy of the estimation of willingness by cross-validation, and found that the method correctly estimated high and low willingness with an accuracy of up to 72.8%.

Next, we introduced an adaptive dialogue strategy using this willingness recognition model and implemented it in a dialogue robot. In the adaptive dialogue strategy, if the estimated willingness is high, the robot continues the topic of the previous question, and if the estimated willingness is low, the robot switches the topic. Question selection by topic continuation/change was realized by exploring a pre-constructed question graph (a tree graph in which questions are arranged based on topic relevance). A dialogue experiment was conducted with 27 participants to evaluate the effect of adaptive dialogue strategy. The dialogue experiments were compared between the proposed adaptive dialogue strategy system and a random strategy system with random topic continuation/transition. The experimental results confirmed that the adaptive dialogue strategy gave users the impression that they were listening with more interest, and also significantly increased the number of

utterances with high willingness. This showed that even with less-than-perfect estimation accuracy, it is possible to motivate users to speak through adaptive dialogue strategy.

In order to improve the accuracy of multimodal inner state estimation and to analyze the accuracy of attitude estimation due to individual differences in multimodal features and the sensing environment, this study worked on refining the attitude estimation using the pre-built external corpus Hazumi1911. We trained and evaluated a model that added biodata and facial landmark features in addition to prosody and posture features. As a result, the addition of the features used improved the accuracy, and the individual differences in estimation accuracy decreased. We evaluated the accuracy of models trained on two different corpora with different sensing environments. The models were trained on the Hazumi1911 corpus and evaluated on the accuracy on data outside the corpus (a newly collected corpus of interview dialogues). The results showed that the accuracy of the models on data outside the corpus decreased, and the difference in accuracy between individuals also increased. However, the decrease in accuracy was smaller for models that used more features, and the difference in accuracy between individuals was also smaller.

A question generation method based on a large-scale language model (LLM) was proposed for the purpose of making adaptive dialogue strategy applicable to arbitrary topics. We implemented an improved interview robot system that incorporates updated willingness recognition model with extended features used and LLM-based adaptive question generation. The results of a dialogue experiment with 30 interviewees showed that the degree of self-disclosure of the interviewees improved when the adaptive strategy was used compared to the random strategy.

In summary, this thesis presents the results of an analysis of multimodal inner state estimation based on nonverbal information during dialogue, the implementation of an adaptive dialogue system, and its impact on dialogue. Our results show that an adaptive dialogue strategy increases user willingness, promotes self-disclosure, and lead to better interviews even with user adaptation using imperfect inner state estimation models.

These results will lead to new applications of dialogue technology through interview techniques that promote self-disclosure of the subject and elicit deeper narratives. For example, by eliciting the user's unspoken feelings and inner narratives, it will be possible to improve counseling and service recommendations. The realization of such assistant technology that empathizes with the user's inner world and proactively suggests solutions to the user's problems will greatly improve future human-computer interaction.

Keywords: Sentiment Analysis; Physiological Signal Processing; Machine Learning; Multimodal Signal Processing; Dialogue System.

Acknowledgements

First of all, I would like to express my sincere gratitude to Professor Shogo Okada at the Japan Advanced Institute of Science and Technology for his warm guidance and encouragement throughout the course of this research. I am grateful to Professor Okada for his guidance and support over the past eight years since I was a master's student, which was also the starting point of this research.

Associate Professor Kiyoaki Shirai of the Japan Advanced Institute of Science and Technology provided me with important feedback on my paper and also supervised my small-scale research project in the field of natural language processing. I am deeply grateful.

Professor Kokolo Ikeda of the Japan Advanced Institute of Science and Technology, Professor Kazunori Miyata of the Japan Advanced Institute of Science and Technology, and Professor Yukiko Nakano of Seikei University provided important reviews of my paper. I would like to express my sincere gratitude to them.

Finally, I thank all of my colleagues at Okada ' s laboratory (Tokyo and Ishikawa), and my family, and my friends for their support.

Contents

List of Figures

List of Tables

Chapter 1

General Introduction

Human-Computer Interaction (HCI), Human-Robot Interaction (HRI), and Human-Agent Interaction (HAI) are fields that focus on interactions between humans and machines. In this domain, various methods have been developed for users to convey information to computers, ranging from Command User Interfaces (CUI) to Graphical User Interfaces (GUI), and even hardware devices like touch panels and voice input systems . These technologies enable people, including those with physical disabilities, to communicate effectively. For instance, individuals with motor impairments can convey their intentions using tools like touch panels or voice input systems[1, 2].

For computers to be universally beneficial, they must offer a smooth means of information exchange for all users, regardless of physical or cognitive limitations. To achieve this, computers must be capable of empathizing with the user's unspoken emotions and inner thoughts, estimating what the user wants, and determining the most suitable means to meet that need.

Voice assistants, such as Siri and Alexa, have become popular as human-like conversational interfaces. These voice-activated systems, equipped with both voice input and synthesized speech output, allow users to control home appliances via simple spoken commands[3]. Such systems have opened up new opportunities for individuals who previously faced difficulties in using computers, thus helping to bridge the digital divide.

However, these systems primarily function by linking spoken commands to actions, without deeper analysis or understanding of the user's intent [4]. Users must carefully formulate their questions to obtain useful responses, and the system offers limited assistance in shaping those queries [5].

Advanced dialogue systems provide more sophisticated interactions by supporting conversational exchanges and question-and-answer sessions. These systems are broadly classified into two categories: task-oriented and non-task-oriented. The former supports specific user tasks, such as product searches or reservation-making, while the latter engages users in general, open-ended conversations. The latter are often used in casual "chatting" scenarios and present several challenges in generating appropriate and contextually meaningful responses [6].

Task-oriented dialog systems are implemented individually for each task, and the user can select from these tasks the one that best suits his/her purpose, but this requires the user to have selected in advance what task he/she is interested in. Technology is needed to extract what tasks are available from the vast number of choices and what tasks the user is interested in through dialogue[7].

Large-scale language models (LLMs), such as GPT[8], had a significant impact on the study of dialogue systems. Recursive text prediction and large-scale models trained on large numbers of documents on the Internet have made the system's speech more natural and sophisticated, allowing users to interact with the system using natural language as if they were conversing with a human-like intelligence.

LLMs are tuned by inputs called prompts to behave generally in accordance with the prompts[9], but the content they generate is unpredictable in principle. While this unpredictability makes interactions with LLMs more engaging and human-like, there is also a risk that this unpredictability can lead to harmful empathy toward the user, loss of self-determination, and the generation of content that encourages users to self-harm or commit suicide [10, 11]. Therefore, in order to achieve a system that demonstrates empathy and familiarity with psychological aspects, it is necessary to have a system that is not a complete black box, but a system that empathizes with the user based on appropriate behavioral modeling that utilizes psychological knowledge.

To realize such empathy technology, it is necessary for the system to recognize the user's internal states, such as feelings and attitudes, and to adapt its behavior to them.

With the recent development of sensing technology, many techniques have been studied to estimate human emotions and other internal states from mechanically measurable information [12, 7, 13, 14]. In particular, multimodal behaviors such as posture, facial expressions, and tone of voice that are observed during a conversation and affect the interlocutor are called "Social Signals"[15] and are the subject of active research on technologies to recognize and process Social Signals. The technologies to recognize and process Social Signals are being actively studied[12, 16, 17]. On the other hand, inner state recognition is a technology based on estimation, and the methods available for dialogue systems, especially those in which large-scale measurement such as EEG measurement is difficult, have low accuracy[18]. It is necessary to build a system that can function well even in such cases.

here has also been active research on user adaptation of systems [19, 16]. These adapt the system's interaction behavior based on the results of inner state recognition to promote changes in user behavior and attitudes and accomplish specific system tasks, such as stress relief in interview dialogue or training in interviewing skills. This study set the system task of "eliciting user interest and concern," i.e., eliciting user self-disclosure, and propose a behavioral strategy for the system suitable for realizing this system task.

It has also been suggested that the optimal presentation behavior differs between humans and robots[20], and it is necessary to actually design and evaluate the optimal robot behavior for the task to be achieved.

One means of eliciting information from a person through dialogue is an interview. Interviews are conducted in a individual dialogue format, with one person asking questions and the other responding[21].

In order to improve the quality of information obtained from interviews, theories and techniques for conducting interviews have been proposed. These include techniques for the interviewer's attitude during the interview[21] and methods for structuring questions during the interview[22]. Techniques for attitude during the interview include the interviewer's way of speaking, gestures, and attitude expression. These help to relax the interviewee and improve their willingness to participate in the dialogue and engagement.

One technique that is widely used for structuring questions is the structured interview[22]. This is a method in which the questions to be asked are decided in advance, and is suitable for collecting and comparing information from multiple interviewees[23]. In contrast, an interview in which the questions are not decided in advance is called an unstructured interview. In an unstructured interview, the questions are decided dynamically, so it is difficult to compare between interviewees, but it is suitable for eliciting in-depth narratives from interviewees.

In unstructured interviews, it is important to follow up on topics appropriately, as the next question is decided based on the answers to the interview. Appropriate topic follow-up can increase the willingness of the interviewee to share information, but inappropriate topics can decrease this willingness[21]. Therefore, in order to conduct a successful in-depth interview, it is essential to assess whether the interviewee has the motivation to share information on the topic.

Therefore, this study implements a system with inner state estimation and adaptive dialogue functions, and through dialogue experiments, clarifies whether the system's behavior adapted to the user's inner state encourages the user to reveal his/her inner state, and whether the less-than-perfect inner state estimation works effectively in such applications.

The ultimate goal of this research is to develop an interactive robot that will encourage the user to actively talk about his or her feelings and experiences. If the user is able to verbalize his or her concerns and needs through dialogue with the robot, the system can suggest other systems that are appropriate and enrich the user's life accordingly.

As an exploration toward this goal, this dissertation consists of the following research elements. Tasks 1-3 each contain issues that are developmental improvements (Tasks 1-a through 3-a).

To achieve this goal, this research will consist of two stages (rounds 1 and 2) of experiments and verification.

In the first round of research, we will build a multimodal attitude estimation model and construct an interview dialogue robot system that can adaptively continue or change the topic of conversation, and then verify whether the adaptive interview dialogue we propose increases the interviewee's willingness.

Based on the findings from the first round of results, we extracted new issues and worked on them. The extracted issues were a detailed analysis of multimodal attitude estimation, real-time generation of adaptive questions, and evaluation of the promotion of self-disclosure through adaptive dialogue. In order to verify these issues, the second round of research improved the question generation and estimation models of the dialogue robot, and evaluated the change in estimation accuracy for the feature set of multimodal attitude estimation and out-of-corpus data, analyzed the effects of real-time question generation using LLM and the impact of generation failure, and evaluated the impact of adaptive dialogue strategies on interviewees' self-disclosure, as well as analyzing the internal state of the interviewees to be adapted to and the effects of adaptive dialogue.

With this in mind, this thesis is composed of the following research elements. The research elements are composed of Task 1, 2, and 3, as well as Task 1-a, Task 2-a, and Task 3-a. The elements that were implemented in Round 2 are given the suffix (-a) as advanced research topics.

Task 1. Multimodal willingness recognition

> We construct an online willingness recognition model that estimates willingness by machine learning using multimodal information such as the user's posture and speech utterances during a dialogue.

Task 1-a. Improvement of willingness recognition model

> In order to improve the accuracy of the speech willingness recognition model, we train a model with more features and evaluate the change in accuracy. We also evaluate the change in accuracy under different sensing conditions, assuming that the model will be applied to actual interactive robots.

Task2. Adaptive dialog strategy based on estimated willingness

We implement an adaptive dialogue strategy that adaptively continues/changes topics based on the results of willingness estimation, and an interview robot dialogue system that conducts interviews based on the adaptive dialogue strategy. Through dialogue experiments, we evaluate the impression that the adaptive interviewing robot dialogue system makes on the interviewee.

Task 2-a. LLM-based adaptive question generation

To remove restrictions on topics that can be handled by adaptive dialogue strategy, we will implement automatic generation of system utterances using a large-scale language model (LLM), and investigate the impact on dialogue when unintended utterances are generated by automatic generation and dialogue breakdown occurs.

Task3. Evaluate the effect of adaptive dialogue strategy

Through the interview dialogue experiment using the adaptive interview dialogue system constructed in Tasks 1 and 2, we will evaluate how the adaptive interview dialogue affects the interviewees' interview behavior, such as their willingness.

Task 3-a. Evaluate the effect on self-disclosure

Through experiments using the adaptive interview robot system constructed in Tasks 1-a and 2-a, we evaluate, using quantitative measures, whether increasing the user's willingness through adaptive dialogue promotes the user's self-disclosure.

The outline of this doctoral thesis is shown in Figure 1.2. The tasks that make up this doctoral thesis and the research questions within each task are shown in Figure 1.3, and the outline of the adaptive interview dialogue robot system realized by this research is shown in Figure 1.1.

In Tasks 1, Task 2, and Task 3, we construct an interview robot dialogue system equipped with a machine learning model that estimates the user's willingness during dialogue and an adaptive dialogue strategy that adaptively continues/changes the topic based on the results of willingness estimation, and conduct dialogue experiments to evaluate the effectiveness of the adaptive dialogue strategy based on willingness estimation through (chapter 3).

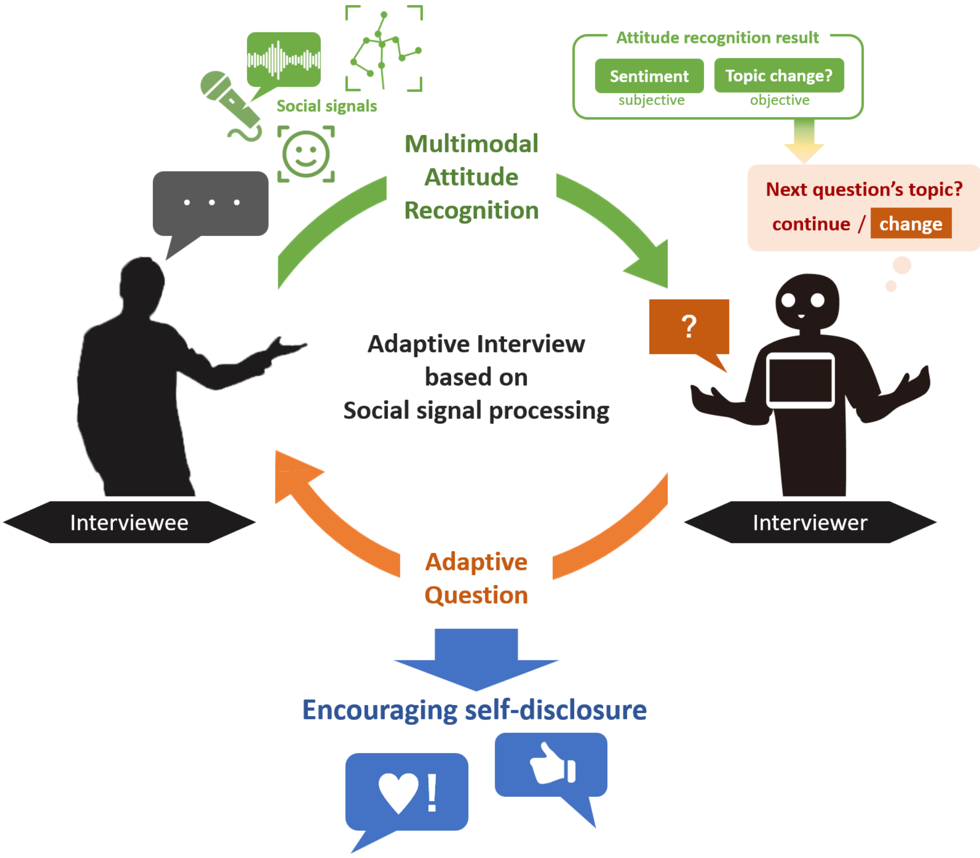

Figure 1.1: Summary of adaptive interview robot system

Based on the evaluation results obtained from Tasks 1, 2, and 3, we worked on Task 1-a, Task 2-a, and Task 3-a as advanced tasks. In Task 1-a, we trained and evaluated a model with more features to improve the accuracy of the attitude estimation model and achieve robust estimation accuracy. In addition to evaluating accuracy using cross-validation, we evaluated how individual differences in estimation accuracy change, and we also evaluated how accuracy changes in environments with different sensing conditions, assuming application to an actual dialogue robot.

Task 2-a is question generation based on a large-scale language model (LLM). We implemented a function to automatically generate the system's question utterances using LLM, and we advanced the system so that it could handle a wide range of topics. We also analyzed the changes in dialogue quality due to the introduction of LLM and the effect on the interviewee's willingness when the question generation fails.

Finally, in Task 3-a, we used the interview dialogue robot system constructed in Tasks 1-a and 2-a to evaluate whether self-disclosure of adaptive dialogue strategies was promoted using psychological scales.

Chapter 2 describes related research surrounding this study, Chapters 3 through 4 discuss each issue, and Chapter 6 concludes this study.

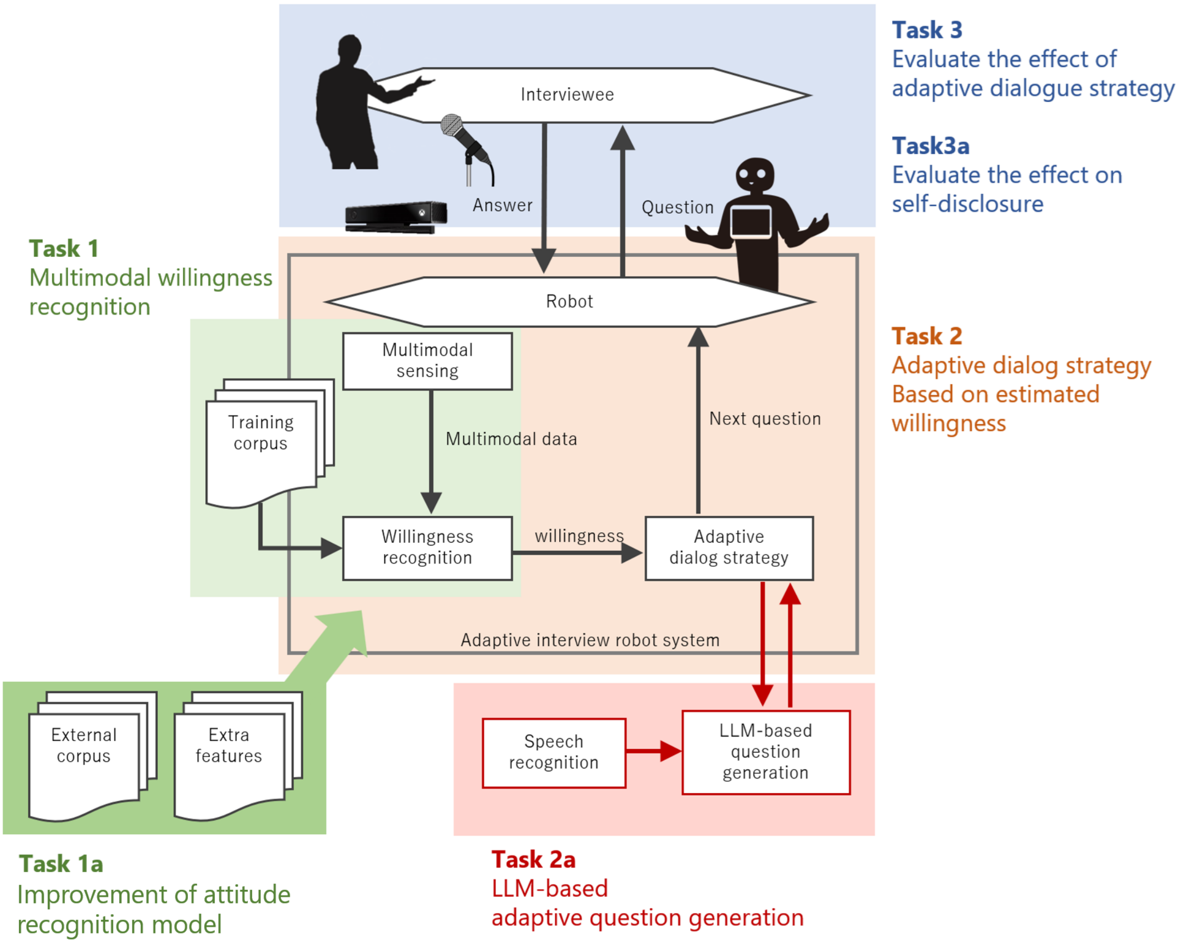

Figure 1.2: Summary of this study

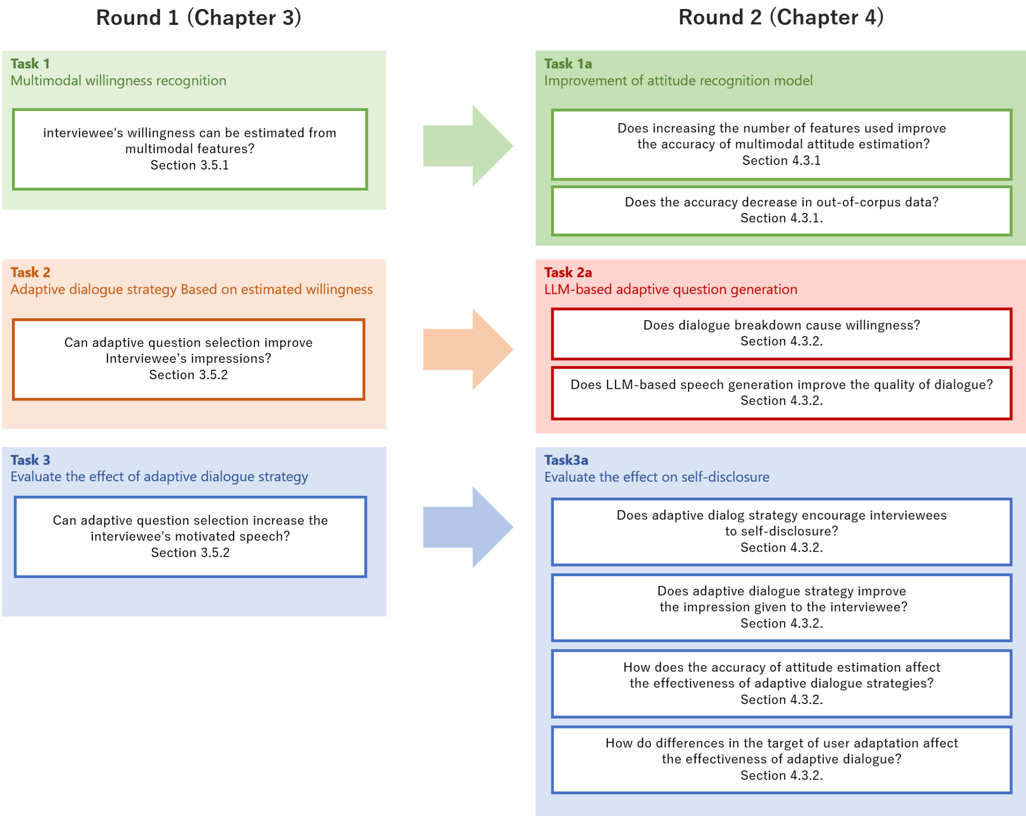

Figure 1.3: Summary of research questions

Chapter 2

Related Works

In this chapter, we will explain related research on this study. The research area and issues related to dialogue systems, which are computer systems that interact with people, are described in Section 2.1, and the related works on inner state estimation and user adaptation, which are closely related to this study, as well as social signal processing and HRI, are described in 2.2. In addition, related research on evaluation methods for research using robots such as this study is described in 2.4.

2.1 Dialogue system

This research is positioned as a field of dialogue systems. Computer systems that behave like excellent assistants have been dreamed up in science fiction and other fictions, and have been the subject of various important research studies[6].

Dialogue systems, which began with early rule-based chatbots like ELIZA[24], have evolved significantly in recent years due to advancements in multimedia processing technology and the widespread adoption of computers[6]. The dialogue system is now used in a variety of situations, including voice recognition assistants such as Siri and Alexa[25]. Since the dialogue system can be used even if you are not familiar with computer operations, the system can provide services to a wide range of users [26].

The article discusses the benefits of providing services through dialogue systems, which are suitable for users who are not familiar with computer operations. It is possible to provide services that can be used by a wide range of users. In dialogue systems, it is important to collect the information necessary to provide services in an appropriate manner while achieving natural conversation, and various methods have been proposed to achieve natural conversation.

Tanaka et al.[27] developed a system that can respond appropriately and with consideration even when users express their requests in an ambiguous way. Their system classifies ambiguous requests into appropriate system actions, and then infers " what the user wants " from the user's ambiguous statements by looking at the system's assumed tasks. The dialogue system we are aiming for is not limited to the system's assumed tasks. It also deepens the topic through repeated interactive dialogue, and encourages the user to disclose what they want.

Among these techniques, the generation of system utterances using generative models such as GPT has attracted particular attention in recent years. In particular, since the advent of LLM such as GPT-3[9] that can adjust its behavior according to prompts, various dialogue systems using LLM have been studied [28, 29].

In these systems, the LLM-based dialogue agent engages in continuous conversation based on a single prompt, with user-adaptive behavior driven by the LLM. These methods can achieve a high degree of user adaptive behavior due to the flexibility of the output from the LLM. However, because they are based on probabilistic models, there is a risk that the system's output may contain errors.[30]. The interview dialogue robot system developed in this study utilizes LLMs as one method for generating system utterances, while user-adaptive speech behavior is controlled by a separately defined algorithm. We also evaluate the impact of unintended speech production by the LLM on the dialogue.

Recent advances in multimedia processing technology have enabled dialogue systems to process multiple modalities beyond just language.

Modality is the form of input/output used by the system to present speech, which can be text input, voice, credentials, and many others. A dialogue system that handles multiple such modalities simultaneously is a multimodal dialogue system.

In human communication, nonverbal cues such as voice tone and facial expressions are crucial for understanding the emotions and attitudes of others[31]. In particular, signals that convey attitudes and emotions to the interlocutor through nonverbal information such as posture and tone of voice are called "social signals," and the field that deals with these signals is called " Social Signal Processing (SSP). Previous research in this field is described in detail in 2.2.

The analysis not only covers the modalities that dialogue systems receive and analyze, but also the modalities in which the system presents information to the user. The forms of interface through which the system interacts with the user are diverse, and include not only text input on the screen, but also androids, smart speakers, and virtual agents that animate on the screen as characters.

Among these, many researchers have investigated “dialogue robot systems”, in which the dialogue system uses a robot as a device for representation.

Spiliotopoulos et al.[26] selected an interactive robot as an assistant robot for hospitals. They mentioned the advantage that interactive robots are suitable for users unfamiliar with computer operations because they do not require input device operations such as a mouse or keyboard.

Schneider et al. [32] explored the design of an assistant system to support dieting and rehabilitation. They compared a robot that provides support and assistance for dieting and other activities, a virtual agent, and a human assistant through user interaction experiments. They focused on changes in the user's exercise time and investigated which of the robot and virtual agent could produce a performance closer to that of a human assistant. The results showed that users tended to exercise for the longest time when paired with a robot and had a high degree of liking for the assistant.

Li[33] explored the effects of physical embodiment and physical presence through a literature review of studies comparing the impressions of robots and virtual agents on users. The results showed that the same robot or virtual character is more convincing and more positively perceived by users when physically present in front of them than when presented as an videofeeds.

These studies suggest that systems designed to closely interact with or advise users benefit from the presence of an entity, such as a robot, behaving as if it were engaged in a human conversation.

In this research, we have placed a robot that acts as an interface for dialogue in a dialogue system. While making use of the advantages of robots, this research focuses on the development and analysis of dialogue behavior. In other words, the results of this research can also be applied to systems other than robots.

Inoue et al. [34] constructed a listening dialogue system running on the android “ERICA” and compared the evaluation results of the constructed system through dialogue experiments with those of the WoZ method, which simulates the same system. Their system generates speech behaviors, such as back-channeling and detailed questions, based on predefined rules in response to the user's speech, thereby progressing the conversation while showing empathy toward the user's statements. Experimental results showed that the system was as effective as WoZ in terms of basic listening skills such as speaking ease, focusing on the conversation, and active listening. In this study, user adaptation is performed based on the willingness to speak, which is estimated from nonverbal features.

These previous studies [35, 34] analyzed the effect of user adaptation in dialogue using various androids. All of the robots mainly serve as an interface in dialogue, performing speech synthesis functions and expressing gestures in accordance with the content of the conversation. Due to the convenience of using robots that were available at the time of the experiment, this study uses two different robots (Pepper and Sota), both of which are used as a common dialogue interface that performs the same speech synthesis function and simple gesture expression.

Naito et al. [20] conducted an experiment to evaluate the communication strategies that are optimal for human and robot interactive robots performing in-store customer service. The experiment was conducted via an online video survey, and the results showed that direct communication was more suitable for customer service tasks than polite human-imitated communication in the case of robots, and that optimal communication strategies for task performance may differ between humans and robots. This study implement a robot that conducts interviews based on the adaptive dialogue strategy, conduct dialogue experiments, and evaluate whether the adaptive dialogue strategy promotes self-disclosure or not, in order to confirm that the proposed method actually works effectively in a robot.

Clavel et al. [36] organized the research issues in dialogue systems that consider both the social aspects and tasks of chatbots and assistants. They pointed out the importance of integrating social-oriented systems that attract users to the system through appropriate social signal processing and task-oriented systems that perform appropriate dialogue according to the tasks of the system. The adaptive interview dialogue system proposed in this research has two aspects: a social-oriented element that enhances the user's motivation to speak through adaptive dialogue based on social signals, and a task-oriented system that is an interview agent that elicits in-depth narratives by promoting self-disclosure. It can be said to be one form of the integrated social-oriented and task-oriented systems pointed out by Clavel et al[36].

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).

Social signals are signals that are conveyed to the person you are talking to through nonverbal information such as posture and tone of voice, and they have an effect on the other person who receives them[37, 15]. The field that deals with these is called "Social Signal Processing (SSP)". There are two approaches to SSP: one is to estimate the internal state of the other person from the observed social signal, and the other is to achieve the goal by appropriately incorporating social signals into the system output. We will introduce research on estimation in section 2.2.1 and research on system output in section 2.2.2.

2.2.1 Estimation

There are research going on into estimating the internal state of users from their multimodal behavior in interactions between humans and agents or robots. Most previous research has focused on estimating engagement in various communication settings (monologue to an audience, two-person or small group). Engagement is defined as an attitude that determines the quality of interaction in [38]. 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 dialogue[39]. Nakano et al. [40] 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. [39] 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.

Hirano et al. [12] 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 dialogue systems. They enhanced the multitask learning framework utilizing weakly supervised learning (WSL) algorithms for which the target label is not necessarily accurate.

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. [41] 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.

Sidner et al. [42] 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. [41] and Nakano et al. [40] used the definition in [42] to annotate the engagement level in their research.

Oertel et al. [43] classified the definitions of engagement used in related works and concluded that engagement is the attitude observed as a result of interest in dialogue, 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. This study aimed 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.

Komatani et al. [44] created a multimodal dialogue corpus, Hazumi, which includes dialogue between virtual agents and humans. Several studies, (e.g. Katada et al.[14]), have focused on developing multimodal sentiment analysis that use the Hazumi corpus to estimate topic continuity and user sentiment, but the impact of applying the corpus to real-world robot scenarios is underexplored.

2.2.2 System output

In addition to estimating social signals, research is also being conducted into how systems can interact with people. This includes research into how systems can express their own attitudes and into agent systems that can adapt to users' social signals.

Virtual agents with social signal sensing have recently been developed for communication skills training; Mohammed et al.[16] developed the dialogue 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.

Tanaka et al. developed a dialogue system that teaches social communication skills through dialogue with people with autism spectrum disorder (ASD) [17] 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 [45] and developed a computer avatar with spoken dialogue to observe the communication behavior of participants with dementia [46].

Several studies have focused on the detection of user interests and concerns. Hirayama et al.[7] 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 .

Chiba et al.[13] estimated the user's level of interest in the dialogue 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.[47] created a corpus of dialogue data for the study of dialogue and user interest.

Tomomasu et al.[48] proposed a method to determine whether a user is interested in a particular topic using facial expression recognition and prosodic information of speech utterances .

Batrinca et al.[49] 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).

Weber et al. [50] 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. [51] presented dialogue 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 dialogue 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. [52] 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. [19] 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 [53]. In particular, the system in [19] 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 post-traumatic 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.

Jeong et al.[54] developed a robot psychology coach that was deployed in dormitories and interacted with university students. They deployed the robot psychology coach in the dormitory and showed that long-term intervention and personalization between university students and robots contributes to improving mental health. Their robot implements simple psychological interventions and interactions such as games over the long term. This study focuses on real-time attitude estimation and behavioral change during dialogue.

2.3 Interview dialogue system

2.3.1 Interview Theory

Interviews have been used as a means of extracting information from the subject through dialogue in various situations, and various interview theories have been studied.

Interviewer techniques can be broadly divided into those related to the interviewer's various behaviors and those related to the content of the interview itself.

The former is due to the interviewer's various attitudes and behaviors, such as the way they speak, sit, look, react, and interrupt, and it can give a better impression to the interviewee, help them concentrate on the interview, and encourage them to participate in the interview [21].

The latter is mainly about how to structure the questions you ask your interviewees. For example, structuring an interview means fixing the questions you will ask in advance, and depending on the degree of this fixation, interviews are classified as structured, semi-structured, or unstructured interviews[22]. Structured interviews are suitable for comparing multiple candidates in recruitment interviews and for gauging the abilities of interviewees[23]. On the other hand, because the questions are predetermined, they are not suitable for gathering information beyond that which has been planned in advance. In an unstructured interview, only the theme of the questions is decided, and the questions are decided dynamically based on the interviewee's answers. Because the interviewee has the initiative in the topic, if the interviewer has information they want to ask about, it is difficult to control the topic, and the interviewer's ability is put to the test, but on the other hand, with flexible questions, it is possible to draw out deep narratives from the interviewee that could not be imagined in the preliminary stages. This study focuses on unstructured interviews in light of the long-term goal of " drawing out what users want to do and what they are interested in " . Unstructured interviews make it possible to collect a wide range of stories from interviewees through flexible content. This study will look at one of the techniques used in this process, topic follow-up.

2.3.2 Self-disclosure in dialogue systems

Interviews are a method of eliciting information from the subject through dialogue, and eliciting the subject's opinions, interests, and other personality traits, in other words, encouraging self-disclosure, is an important technique.

Self-disclosure refers to the act of revealing oneself to others in order to let them know what kind of person you are. According to Social Penetration Theory by Altman et al. [55], there are multiple stages of self-disclosure, and deeper self-disclosure involves negative content such as one's own weaknesses or socially undesirable content. Disclosing one's own deep opinions is an act that carries a great risk of being criticized by the other party, and such self-disclosure is less likely to occur when one's intimacy with the other party is low.

Soleymani et al. [56] 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 processes and sentence construction were important indicators of intimate self-disclosure. Head gestures such as nods and speech pauses were also associated with self-disclosure. This research not only estimates and analyzes the level of self-disclosure, but also focuses on whether self-disclosure is promoted by the system's dialogue strategy.

Mitsuno et al. [57] proposed a chat dialogue agent that interacts with users over a long period of time and gradually moves on to topics that require deep self-disclosure. Their 10-day experiment showed that the agent that gradually moved on to deep topics increased the degree of self-disclosure by users compared to the agent that did not do so. Their agents deepen the topic over time regardless of whether the user responds or not, but our method aims to promote self-disclosure in a short period of time while also controlling the topic adaptively in response to the user's responses.

Niwa et al[58]. proposed a self-disclosure evaluation scale. Based on the social penetration theory proposed by Altman et al.[55]. Niwa et al. created this index as a tool for evaluating the degree of self-disclosure between human beings, but it is also used to evaluate the degree of self-disclosure to dialogue robots and agent avatars[57]. The self-disclosure we aim to promote in this study is similar to that dealt with by Niwa et al. from the perspective of " having the interviewee speak their true feelings " . We will use this index as an evaluation index for our proposed method.

2.3.3 Interview Dialogue System

There are many studies on dialogue systems that conduct interviews.

Mohamed and his colleagues developed a virtual interview agent called "MACH" that detects social signals. "MACH" was developed to help students practice their communication skills for 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 [16].

Saito et al. [52] 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. [19] 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 [53]. In particular, the system in [19] 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 post-traumatic 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.

Kobori et al. [59] developed a text-based interview dialogue 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 dialogue. Our research focuses on the changes in the interviewee's willingness that occur as a result of choosing the content of the dialogue itself.

Inoue et al. [39] propose to generate more in-depth questions by analyzing words in interviewees' responses. This study test the effectiveness of using LLMs, recorded dialogues, and dialogue transcripts to generate follow-up questions related not only to words but also to context.

Chiba et al. [13] investigated a method for estimating the interviewee's willingness to continue the dialogue from multimodal features, with the goal of making the interview dialogue 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 dialogue conducted by human interviewers, but we consider the change in interviewee's willingness when using a robot as an interviewer. Ishihara et al. [60] 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.

2.4 Evaluation methods of dialogue robot

In order to research dialogue robots and interview dialogue agents, it is necessary to formulate evaluation metrics to assess their effectiveness. Here, we will introduce related research on the evaluation of dialogue systems and the evaluation of robots.

2.4.1 Evaluation of Dialogue

With regard to the linguistic performance evaluation of dialogue systems, there are mainly two types of evaluation: subjective evaluation by the person who interacted with the system or a third party who observed the interaction, and objective evaluation such as keyword counting of linguistic behavior and content of statements.

In terms of subjective evaluation, "context relevance and coherence" as used in the study by Niu et al. There are also competitions based on common tasks, where a common evaluation scale is defined to compare multiple systems. These include "How do you feel about speaking with this socialbot again?" in the Amazon Alexa Prize[61] and "fluency" in the convAI2[62] challenge.

Higashinaka[63] compares evaluation metrics for various dialogue systems, and points out that many evaluation items are heavily dependent on the evaluator's subjective judgment, making it difficult to compare them as engineering research. In addition, Inaba[64] compared and analyzed the pros and cons of manual and automatic evaluation methods for evaluating dialogue systems, and concluded that it is important to combine both methods.

This study use both the impression evaluation by the test subjects themselves and objective quantitative indicators such as the ratio of motivated speech by the interviewees as impression evaluations of the proposed interview dialogue system.

2.4.2 Evaluation of robots

In research in the HAI/HRI field, a wide variety of virtual agents and robots are used. A variety of robots are used, including relatively small robots such as NAO and Sota, larger robots that resemble humans such as Pepper and ERICA, and non-humanoid robots.

Which robots are used in research depends not only on the design of the robots themselves, but also on various factors such as equipment procurement and the location of the experiment, but it is also important to compare the various research projects that use a wide range of robots and integrate them into a larger research trend. As a common evaluation metric for these robots, Godspeed was proposed by Bartneck et al.[65] and has been adopted in several studies[66]. This has made it possible to evaluate various robot research in a cross-sectional manner.

Regarding comparative studies of robots using Godspeed, there is a survey by Rossi et al.[67] that evaluated the differences in impressions when Pepper and NAO performed the same task, and a survey by Martina et al.[68] that summarized the human-likeness of robots and the impressions of observers.

This study conducted impression evaluation using unique items based on the task of the interview robot, and also conducted evaluation using Godspeed, so that we could compare these studies. In addition, the two experiments conducted in this study used different robots (Pepper and Sota). In the experiments, impression evaluation was conducted between systems that switched dialogue strategies, but the evaluation was conducted between the same robots, and care was taken to prevent the appearance of the robots from having an effect. In addition, when evaluating the differences between the two experiments, only the dialogue transcripts were given to the evaluators so that the impression of the robot would not have an effect.

2.5 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. [69] 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 [50], 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 dialogue with the robot (system) through questionnaire surveys. We focus on evaluating not only the impression of users toward the dialogue 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 dialogue strategy based on the estimated willingness level changes the user's behavior, eliciting utterances with high willingness levels and increasing self-disclosure.

Third, we will compare the effectiveness of adaptive question generation by comparing conversations based on a list of prepared questions with conversations based on adaptive question generation by LLM.

Inoue et al.[39] propose a method for generating follow-up questions based on the results of an analysis of the words contained in the answers of the interviewees. This study use LLM, recorded dialogue, and dialogue records to generate follow-up questions that are not only related to words but also to the context. Inaba[64] pointed out that in the impression evaluation of dialogue systems, human subjective evaluation may give a higher impression to systems that perform advanced speech generation than to systems that perform pre-programmed speech generation. This study asked pre-programmed questions in Research Round 1 and perform advanced speech generation (question generation using LLM) in Round 2, and compare the results of each.

Fourth, we will examine whether adaptive interview dialogue promotes self-disclosure by the interviewee. Mitsuno et al. [57] showed that a dialogue strategy that gradually deepens the topic can increase users' self-disclosure.

The adaptive interview dialogue strategy proposed in this study also deepens topics in stages, but in contrast to the system proposed by Mitsuno et al.[57], which deepens topics at a fixed pace over a long period of dialogue with the user, our system focuses on promoting self-disclosure in a single short dialogue by performing real-time user adaptation based on social signal processing.

Chapter 3

Developing of interview dialogue robot system based on multimodal attitude estimation

3.1 Introduction

Recent developments in nonverbal behavior recognition techniques enable systems to recognize social signals and social behavior [37], 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 [41] and in user engagement estimation for adapting the dialog strategy [40]. 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 [21], 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 [21].

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. [56] 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[21].

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 [21] 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 chapter 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 chapter is organized as follows. Section 3.2 presents the interview robot system with the speaking willingness recognition model. Section 3.3 discusses the multimodal interview corpus used to train the willingness recognition models. Section 3.4 describes how the speaker willingness recognition model is trained based on multimodal features. The experimental setting for evaluating the system is described in Section 3.5, and The results are presented in Section 3.6. Finally, the results are discussed in Section 3.7, and The research is concluded in Section 3.8.

3.2 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 Figure 3.1.

This section describes the humanoid conversation robot (Section 3.2), the sensing environment for the interview robot system (Section 3.2.1) and the interview interaction scenario and adaptive question selection based on the willingness recognition results (Section 3.2.2).

The proposed interview robot aims to elicit information from the interviewee through an adaptive question selection strategy. Figure 3.2 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.

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 [70]. 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 the humanoid interview robot module.

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

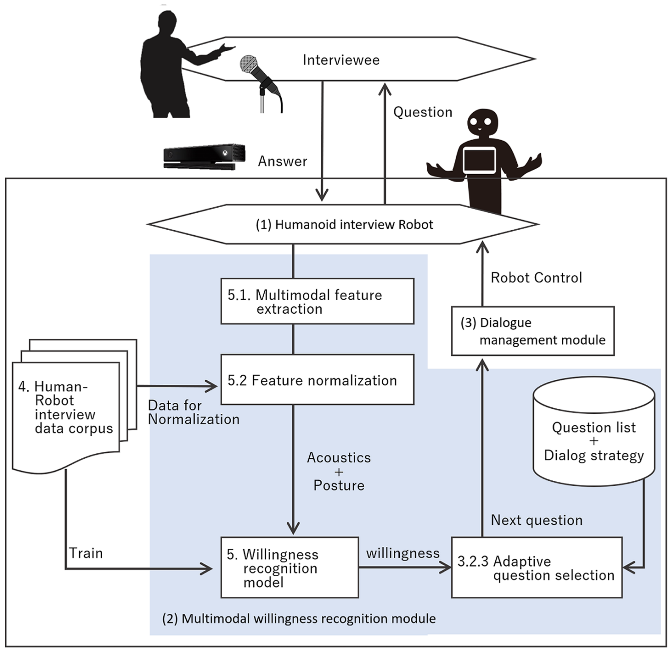

Figure 3.1: Interview robot system

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.

3.2.1 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.3.2). The arrangement of the participants, the robot, and the sensors is shown in Figure 3.3. 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

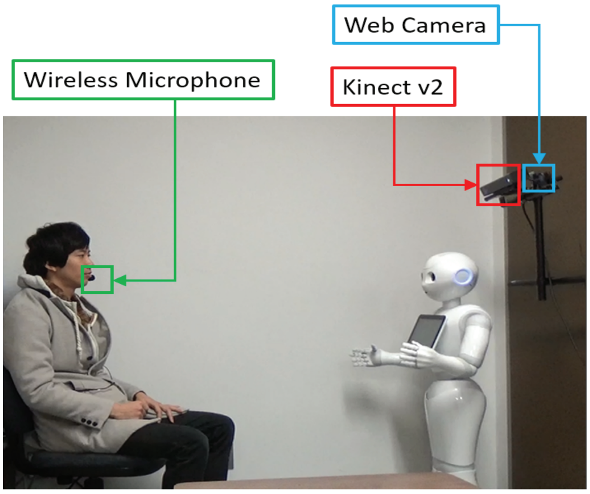

Figure 3.2: Interview scene with the interview robot system

recognition model of willingness are explained in Section 3.4.1.

3.2.2 Adaptive question selection

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 Figure 3.4. 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 Figure 3.4. 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

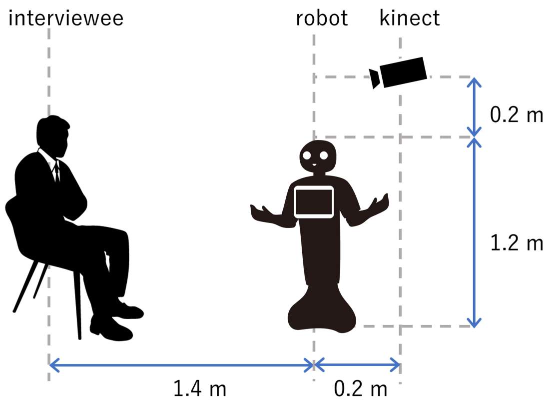

Figure 3.3: Layout of the interviewee, the robot, and the Kinect sensor

by switching the two search methods (depth-first search and breadth-first search)[71] 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.

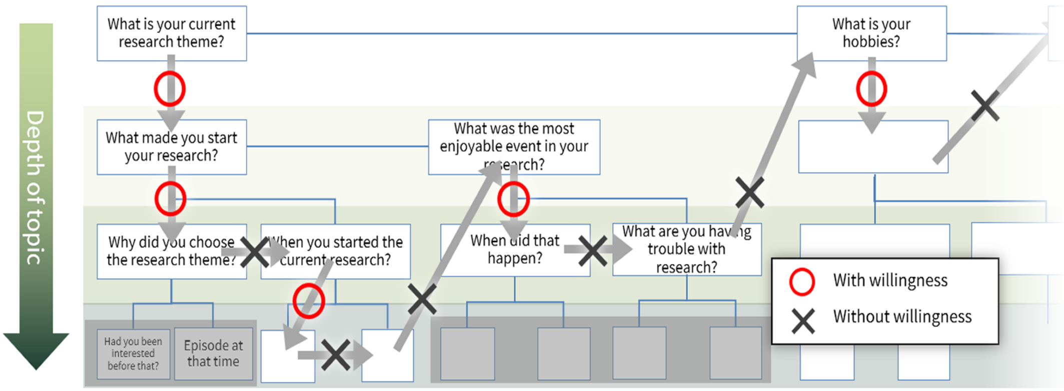

Figure 3.4: Example of adaptive choice based on estimated willingness. Each question node, represented by a box, is arranged in a tree structure. Based on the estimated willingness, the next question is selected from this tree structure.

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.

3.3 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.

3.3.1 Corpus setting

To collect this data corpus, we recruited 8 interviewees (7 male/1 female, aged 22-30 years). The interview robot system asked questions in order based on the prepared list shown in Table 3.1.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.

Table 3.1: 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.

3.3.2 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 Figure 3.5. 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 the observed multimodal features are varied in such a conversation set-

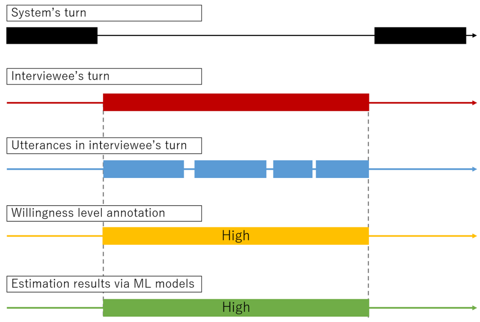

Figure 3.5: The section that performs the robot's turn, the interviewee's turn and the willingness annotation

ting. 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.

3.4 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.

3.4.1 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.

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 [72] 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 $(15 \times 4 + 1(speakinglength))$.

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.

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 different ($p < 0.05$) between the high/low willingness groups are listed in Table 3.2.

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

Table 3.2: 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)

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.

3.4.2 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:

$$x_n(t, d) = \frac{x(t, d) - X_{min}(d)}{X_{max}(d) - X_{min}(d)} \tag{3.1}$$

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 3.1.

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:

$$x_n(k', t', d) = \frac{x'(k', t', d) - X_{min}(k', d)}{\bar{r}},$$
$$\bar{r} = \frac{1}{N_t} \sum_k r(k, d) \quad (3.2)$$

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

3.4.3 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.

Linear support vector machine (SVM)

In the binary classification task, linear SVM models[73] 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.

Random forest

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

3.5 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.

3.5.1 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 3.4.2).

3.5.2 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 3.5.1. 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.

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.[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”.

[2]Research Ethics Committee of the Tokyo Institute of Technology (Application No.A17051) Declaration of Helsinki on Ethical Principles for Medical Research and Ethical Guidelines for Medical and Biological Research Involving Human Subjects by the japanese government.

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 3.3. 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.

Table 3.3: 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?

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 3.5.2, 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.

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 3.6.2. 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 3.7.

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.

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 3.6.2. 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 section3.6.2.

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 3.4: 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

3.6 Results

3.6.1 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. Table3.4 shows the classification accuracy of the willingness estimation models.

Comparison between multimodal features:

In terms of the comparison between the unimodal models (acoustic or posture), Columns 3 and 4 in Table3.4 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 Table3.4 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.

Effect of approximate normalization:

As noted in Section 3.4.2, 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

Table 3.5: 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

(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 3.5 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.

3.6.2 Evaluation of the proposed strategy’s efficiency

In this section, we present the results obtained from the experiments described in Section 3.5.2, which are based on quantitative measures.

Comparison of utterances with high willingness

Table 3.6 shows the number of utterances and the percentage of utterances with high willingness. Column 2 of Table 3.6 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

Table 3.6: 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

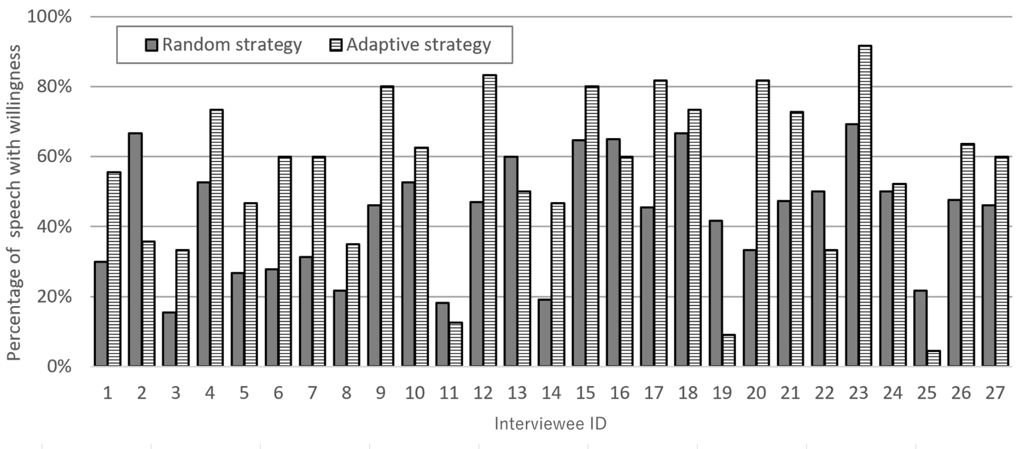

Figure 3.6: Percentage of “with high willingness” utterances per interviewee.

random strategy (43.1%) was used. Conversely, the percentage of exchanges shown in column 3 of Table 3.6 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 Figure 3.6. In the case of the adaptive strategy, the percentage of willingness was higher for 21 of 27 individuals. Figure 3.6 shows that the 21 interviewees tended to speak with high willingness more often when the adaptive strategy was used.

Questionnaire survey for impression of the system

Table 3.7 and Figure 3.7 show the results of the questionnaire conducted in Section 3.5.2.

Table 3.7: 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	

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% 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

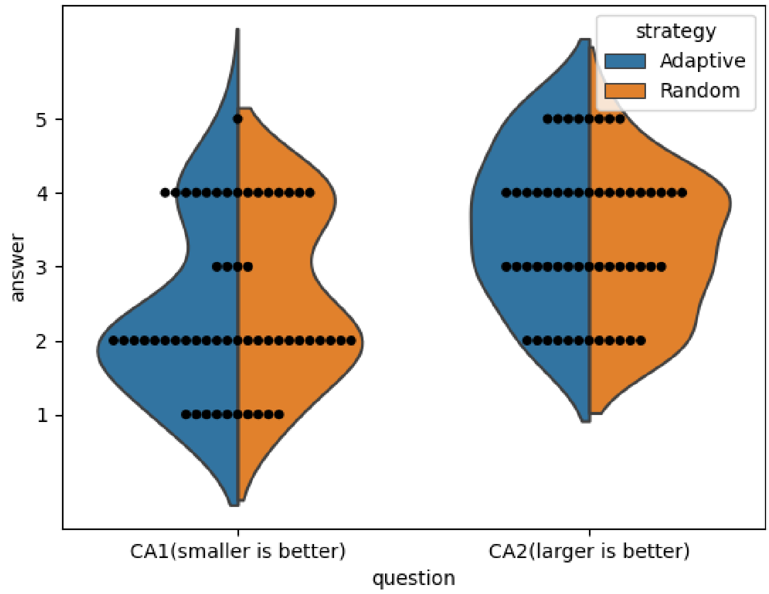

Figure 3.7: Violin plots showing the distribution of respondents for CA1 (responses to CQ1) and CA2 (responses to CQ2) of the questionnaire in the interview experiment. The number of respondents for each option is represented by black dots.

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 3.3 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.

Impressions of adaptive interview dialog

Table 3.8 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 3.9 also shows the number of respondents for each option for the question about whether they felt any strangeness in the dialog or the intensity of the strangeness. In Table 3.9, the number of respondents who answered "no" to question GQ2 is assigned to option "0 (There was no discomfort)".

Table 3.9 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

Table 3.8: 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 3.9: 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

overcoming the challenges of natural language processing described above, accelerating the process of willingness estimation is also an important future work.

Case study

We analyzed the case of interviewees whose percentage of willingness was lower in the case of the adaptive strategy. Figure 3.8 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 Figure 3.8. 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.

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

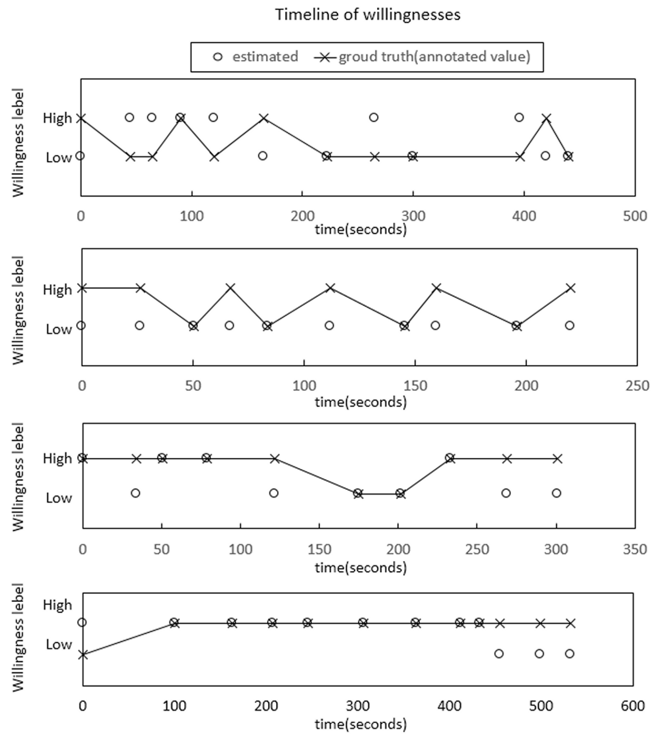

Figure 3.8: The right side of the figure shows the time trends of the estimated and ground-truth willingness; the left side shows the percentage of willingness and responses to the questionnaire (CA1 and CA2). Two examples are shown for each interviewee with high/low percentages of utterances with high willingness: the low group is ID22 and ID16, and the high group is ID9 and ID23.

3.7 Discussion

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

3.7.1 Effectiveness of the adaptive strategy

Tables 3.6 and 3.7 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. [59] 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. [13] 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 [3] 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. [7] 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

[3] The start time when the robot asks questions is controlled by an operator.

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 [7] 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.

3.7.2 Significance of the adaptive interview robot

The advantage of the adaptive interview robot is supported by the findings of [21]. Ben et al. [21] 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. [56] 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.

3.7.3 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.

Accuracy of the willingness recognition model

As shown in Table 3.5, 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 3.6 indicate that following up on topics based on our model increases the percentage 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.

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 3.6 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. [69] 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.

3.8 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).

Chapter 4

Advancement of interview dialogue robot system based on multimodal attitude estimation

4.1 Introduction

The focus of this research is an interview dialogue robot that selects questions adaptively based on the interviewee's (person being interviewed) willingness to talk.

In Chapter chapter:TAC, we developed a system that estimates willingness from the posture and acoustic features of the interviewee's behavior, and then selects next questions based on the estimated results. The results of the experiment showed that the system could estimate willingness with 72% accuracy based on the multimodal features of the interviewee. It was also shown that the adaptive dialogue strategy, which selects questions based on the estimated willingness and adaptively decides whether to continue or change the topic, increases the proportion of motivated speech compared to when the topic is continued or changed at random.

In this chapter, we will develop this fundamental achievement and deepen our research from the following three perspectives:

4.1.1 Improvement of multimodal attitude recognition

For dialogue systems, it is important to be able to demonstrate stable performance when dealing with various systems and various users. To achieve this, it is necessary to have a system that is effective in as many places as possible and for as many people as possible. In particular, in order to have a system that can be applied to such a wide range of situations, it is important that the system can stably estimate the internal state of the interviewee, who is the target of the system.

In order to be used in more places, it is important to be able to maintain high estimation accuracy even when the sensing environment for estimating internal states differs from the one used for training. In addition, in order to be used by more people, it is important to reduce the individual differences in internal state estimation. In machine learning tasks that use multimodal features, individual differences such as body size and personal habits have a significant impact on estimation accuracy, so it is important to reduce the impact of individual differences. If the impact of individual differences cannot be eliminated, the difference between people with high and low estimation accuracy will increase, leading to variations in dialogue quality in real-world applications.

Some previous studies have focused on estimating internal states, and various features have been used[56, 14]. In this chapter, in addition to the postures and prosody used in chapter3, features such as facial landmark features (capturing facial expressions, etc.) and biodata such as heart rate are used as features.

In this chapter, the variation in estimation accuracy for each interviewee is analyzed by analyzing the estimation accuracy when the number of feature values is increased and the accuracy of the attitude model when estimating across two corpora with different sensing environments.

4.1.2 LLM-based adaptive question generation

In the system described in Chapter 3, the system adopted the next question based on rules from a pre-prepared question graph. While this approach is simple and easy to manage, there is a limit to the variation of questions that can be prepared in advance, and there were cases where follow-up on the topic was insufficient.

The experimental results in Section 3.6 suggest that, despite the high willingness of the interviewees, a topic change due to question depletion can lead to a decrease in willingness. In order to overcome these problems, this chapter will use speech recognition and large-scale language models (LLMs) to generate questions in real time. While the restriction on topics that can be handled by automatic question generation is removed, there is a risk that the system-generated speech will not be appropriate as a question. If the same question is repeated over and over again, or if the speech is not a proper question to begin with, it will discourage the interviewee.

Therefore, preliminary experiment was conducted to evaluate and analyze the impact of dialogue breakdowns first. Next, impression the system had on the interviewee were evaluated through interview dialogue experiments, and also evaluated how the content of the conversation changed through the eyes of a third party.

4.1.3 Evaluate the effect on self-disclosure

In chapter3, it was shown that an adaptive dialogue strategy based on estimating the interviewee's internal state increases the percentage of utterances with high-willingness. However, in addition to the "CQ1:attitude of interest", the impression that "CQ2:unpleasant question" was also increased. In the free-response sections of the individual questionnaires, even if negative emotions were being expressed, it was presumed that the willingness was high, and it was suggested that the system was digging deeper into the topic.

In order to achieve the aim of this research, which is to " encourage people to talk more about what they want to talk about " , it was found that it is necessary to evaluate whether people were able to share more of their innermost thoughts, from a perspective other than simply increasing the percentage of motivated speech. In this chapter, the evaluation of an interview dialogue system was examined using the evaluator scale for self-disclosure created by Niwa et al.[58] and the godspeed questionnarie [65], which has been proposed as a standard benchmark for human-robot interaction.

4.1.4 Correlation analysis of attutude estimation accuracy and dialogue effectiveness

The key to an adaptive dialogue system is a high-precision estimation of the user's internal state and an appropriate user adaptation strategy based on the estimated internal state. The model trained in Chapter 3 had an estimation accuracy of 72% in the binary classification task of high/low willingness, which is not necessarily a perfect estimation accuracy. Nevertheless, the percentage of utterances with high-willingness was increased in the adaptive dialogue using the trained internal state estimation model.

While improving the accuracy of inner state estimation is expected to increase the proportion of motivated speech, inner state estimation is generally difficult[5]. In addition, it is not clear to what extent improving the accuracy of multimodal attitude estimation will contribute to improving self-disclosure, which is the aim of this research.

In this chapter, a correlation analysis is conducted between the various evaluation scores and attitude estimation accuracy of each participant collected through dialogue experiments, and it becomes clear which aspects of the effect of adaptive dialogue can be expected to be improved by improving the multimodal attitude estimation model.

4.1.5 User adaptation targets and effects of adaptation

The system in chapter3 performed user adaptation by estimating "willingness" based on third-party annotations as an internal state that serves as a basis for user adaptation, and then adaptively selecting the next question from a pre-constructed question graph based on the estimated willingness. It has been pointed out that the emotions a person is feeling and the attitude they display to the outside world can be different.

In this chapter, we propose two internal states for user adaptation: "one's own impression" of whether the person's emotions are good or bad, and "third-party observation" of whether the user seems to be feeling good or bad emotions from the outside. When applying the former, the system delves into topics that the interviewee feels positively about, and avoids topics that they feel negatively about. This is considered to be a straightforward approach to the goal of enhancing the emotions of the interviewee. On the other hand, it is necessary for the participant to provide their own annotations and evaluations of how they felt, and because people feel things in different ways, there is a problem with it being difficult to obtain annotated data or quantitative evaluations. There is also a possibility that if the participant is unknowingly interested in something, it will be missed.

In the latter case, if the system can determine that a topic should be continued based on external observations, it will delve deeper into that topic. The behavior of humans when performing similar roles would be similar to this. While third-party annotation is easier to collect data from than self-annotation, there is a possibility that the topic will change in a way that is not in line with the person's interests. On the other hand, if the person's unconscious interest and concern are shown in their attitude, it is possible to capture such topics successfully.

In this chapter, systems were constructed and evaluated for these two adaptation targets.

4.2 Improvements to the dialogue robot system

Figure 4.1 shows an overview of the improved version of the interview robot dialogue system developed in this chapter, and Figure 4.2 shows the system in conversation.

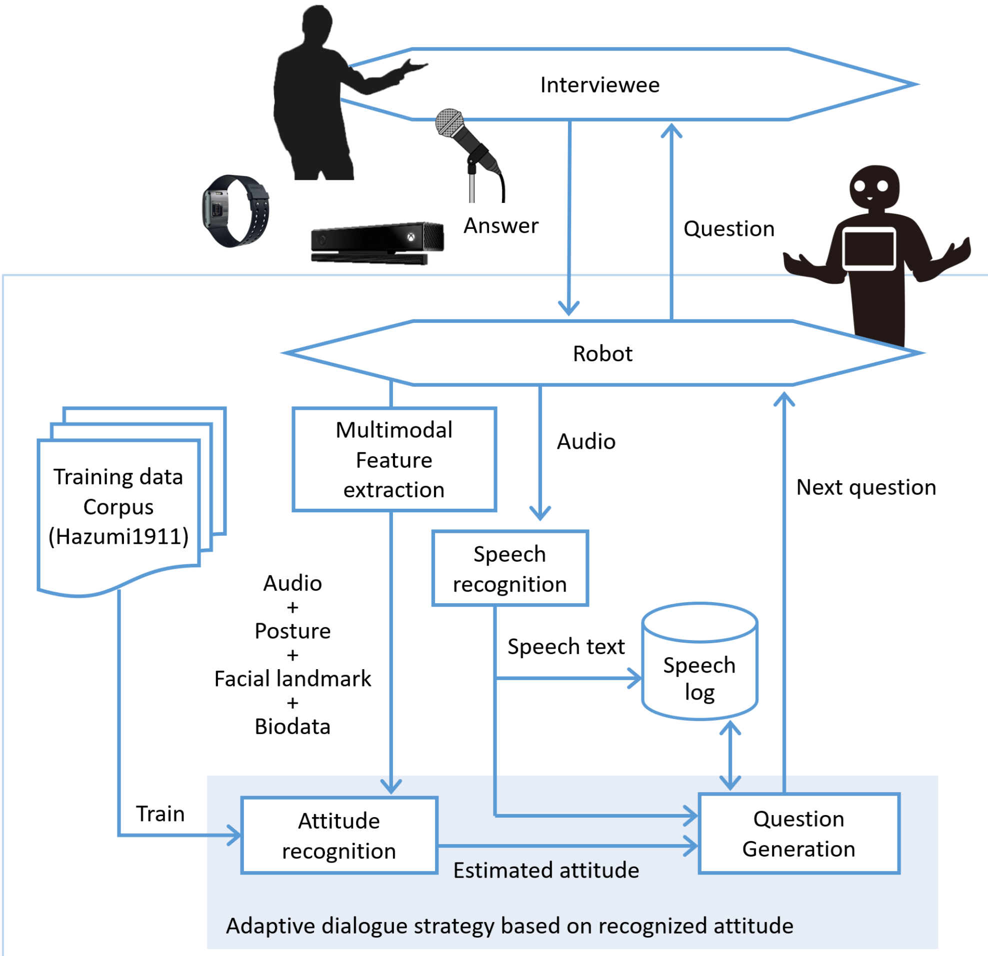

Figure 4.1: Overview of interview robot system(updated version)

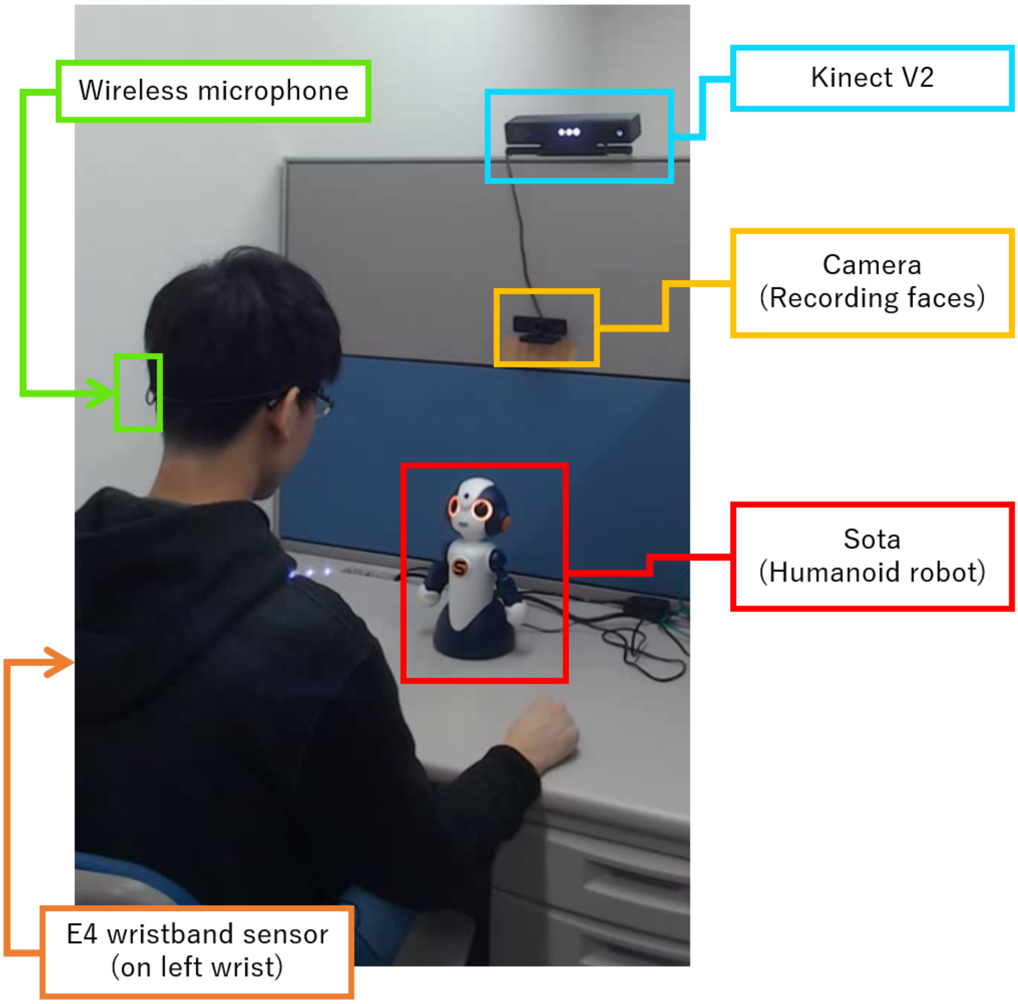

Figure 4.2: Photo of Interview robot system(updated version)

4.2.1 Humanoid conversational robot

The humanoid personal robot Sota was used as the interview robot. Sota is a desktop robot that was developed by VStone. Sota is 280 mm tall and weighs approximately 760 g, and it uses a module that generates speech synthesis and hand and head movements. Sota ' s dialogue is controlled by back-end modules (e.g., a multimodal sensing environment and adaptive question generation).

Figure 4.1 shows an overview of the interview robot system equipped with a Social Signal recognition module. The proposed interview robot aims to elicit information from the interviewee through an adaptive dialogue strategy. The system is composed of three modules: a humanoid conversational robot, a multimodal sensing environment, and adaptive question generation.

4.2.2 Multimodal sensing environment

During the interview, multimodal data were collected via a webcam (Logitech C910, 1080p 30fps), MS Kinect V2, a wearable microphone (Shure PGA31 headset microphone), and a wristband biometric sensor (Empatica E4). During the interviews, the interviewee sat in front of the robot, while the webcam and Kinect sensors were positioned behind and approximately 50 cm above the robot ' s head. The Empatica E4 was worn on the participant ' s left wrist.

In a multimodal sensing environment, postural, acoustic, facial, and biological features were calculated based on the sensing results. The posture features were the joint coordinates estimated by the Kinect sensor, which are the three-dimensional positional coordinate data of 25 joints of the whole body. Acoustic features were calculated from the audio collected by wearable microphones via OpenSMILE [72] and included the mean square frame energy (RMSenergt), mel frequency cepstral coefficient 1-13 (MFCC 1-13), and fundamental frequency (f0). For facial features, the interframe velocity and acceleration of the face were calculated for 12 landmark coordinates around the eyes and mouth extracted from the video camera images using dlib. The biological features included the heart rate and skin conductance (measured by the E4 device). The minimum, maximum, mean, and variance of all these features were computed for each speech segment and used as inputs to the attitude estimation model. Furthermore, the voice data collected by the wearable microphone were transcribed via Whisper for speech recognition and used as input for the question generation module.

4.2.3 Multimodal attitude recognition model

Data Corpus

The Hazumi1911 dataset [44], a multimodal corpus of human-agent dialogue was used. Hazumi1911 contains 2859 exchanges from 30 participants that interacted with an agent in a Wizard-of-Oz setup.

The dataset incorporates diverse data types: posture (3D joint coordinates via MS Kinect), acoustic (prosodic features), facial landmarks, and biometric signals (heart rate, skin conductance). After the experiment, the participants labeled each exchange with self-assigned sentiments and topic continuation preferences.

4.2.4 Two multimodal attitude estimation models

In this chapter, two multimodal attitude estimation models were trained.

Among the annotation labels, "self-sentiment" and "topic consultation" were used in this study. "Self-sentiment" is the participants' own labeling of their feelings (positive/negative) about the content of the dialogue. "Topic continuance" is a third-party annotator's labeling decision on whether the system should continue or change the topic, assuming that the annotator has taken the system's position. In this study, two adaptation strategies, namely self-sentiment and third-party observation, were evaluated. For this purpose, the attitude recognition models were trained using self-sentiment labels as the adaptation target for the self-sentiment strategy and topic continuation labels as the adaptation target for the third-party observation strategy.

In our study, we performed a binary classification task to discern attitude recognition, thus aiding in two-way question selection (topic continuation or switching). We converted the original 8-level Hazumi labels to a binary scale: scores of 5 and above were categorized as 'positive', and scores below 5 were considered 'negative'.

4.2.5 Adaptive question generation

If the system recognized that the topic should be continued based on the results of the attitude recognition for the previous utterance (the response to the previous question), it will continue the topic and ask in-depth questions. If the system determines that the topic should not be continued, it will ask questions that change the topic. This decision is made according to the two models as follows: If the sentiment label estimation result is positive, continue the topic, if negative, change the topic (SS model). If the topic-continue label estimation result is positive, continue the topic, if negative, change the topic (TC model).

The system used LLM to generate questions. During this process, instructions for question generation, examples of appropriate questions based on the user's utterances, and a record of the system's interaction with the interviewee were entered into the GPT model as prompts. To switch question modes, the system adaptively changed the appropriate example questions in the prompts. To accommodate both topic follow-up and topic switching, specific question examples were prepared for each question mode, and the system adaptively switched question examples with each question generation.

In this study, since it was necessary to generate the next question in real time based on the content of the immediately preceding user utterance, Swallow-LLM-7B [74], a GPT model that can rapidly generate Japanese in an offline environment was used.

In the LLM speech generation, the sampling process was disabled and the seed value was fixed. This was done to eliminate randomness as much as possible and to increase reproducibility as much as possible.

4.3 Experimental settings

Using the system constructed in Section 4.2, we evaluated a multimodal attitude estimation model and a dialogue system with an interview dialogue robot that has adaptive question generation based on the estimated attitude.

First, the multimodal attitude estimation model was trained and evaluated. Next, dialogue experiments were conducted using a dialogue system that incorporated the multimodal attitude estimation model. The dialogue experiment was conducted in two parts: a preliminary experiment and a main experiment.

4.3.1 Evaluation of multimodal attitude estimation models

The first experiment was conducted to test two research questions related to multimodal attitude estimation.

The first research question is whether increasing the number of features used for estimation improves the accuracy of multimodal attitude estimation and reduces the difference in accuracy between individuals? We evaluated the accuracy of the model using cross-validation with the Hazumi1911 corpus. The models were compared in terms of accuracy between the model trained on the same feature set as the model trained in chapter 3, P+A (posture + acoustic), and the model trained on the feature set P+A+F+B (posture + acoustic, plus facial landmarks and biodata), which is also available for use with the Hazumi1911 corpus.

The second research question is whether the estimation accuracy of the multimodal attitude estimation model decreases or the individual differences in accuracy become larger for data outside the training corpus? We used a model trained with the topic-continuation labels assigned to the Hazumi1911 dataset as the target variable. We evaluated the change in estimation accuracy due to the data corpus by comparing the estimation accuracy of the model trained on the dialogue experiment corpus collected in this study.

To verify these research questions, we used cross-validation to evaluate the accuracy and individual differences in accuracy. We trained a random forest model in the same way as the chapter3 method, and evaluated the trained model as follows. We used leave-one-person-out cross-validation (LOPOCV) to evaluate the accuracy of the trained intention recognition model.

In LOPOCV, the test data corresponded to the sample observed in the interview session of one participant , and the remaining sample of the other participants was used as the training data. Furthermore, we evaluated the individual differences in attitude estimation accuracy as follows.

In the cross-validation results within the Hazumi1911 corpus, the variances of the accuracies calculated for each one-person-leave-out condition and the accuracies for each individual in the out-corpus condition were obtained and compared. Since variance is an indicator of the variation in values within a set, if the variance is small, it means that the variation in estimation accuracy between individuals is small.

4.3.2 Evaluation of self-disclosure through adaptive question generation and two types of user adaptation

The purpose of the second experiment was to test several research questions regarding adaptive interview (question generation) strategies based on attitude estimation.

The following research questions were tested in this experiment: Does LLM-based question generation lead to more natural interview dialogue? Does question generation improve the impression given to the interviewee? And does an adaptive dialogue strategy promote self-disclosure by the interviewee?

We constructed a dialogue system that incorporated the attitude model trained in section 4.3.1, and conducted interview dialogue sessions using the constructed dialogue system.

Three interview sessions were conducted for each participant. Two of these were topic control sessions using adaptive strategies, and the other was a random topic control session. The topic control sessions using adaptive strategies were conducted using two different attitude estimation models: System (I) used the “Self-Sentiment” estimation model, and System (II) used the “Topic-Continue” estimation model. In the random topic control sessions, System (III) was used. This had the same dialogue flow as Systems (I) and (II), but the system did not perform inner state estimation, and instead randomly decided whether to change or continue the topic for each question.

After the dialogue experiment, the participants were asked to fill in a questionnaire and reflect on the dialogue content, and then to annotate it.

The detailed experimental setup is as follows:

Participants

Ten participants, aged 23 to 63 (average age 44), were recruited through a Japanese recruitment agency, representing a broad demographic of the Japanese population. They received a fixed payment for their participation. Prior to the experiment, participants were informed of their right to withdraw at any time. The experiment ensured minimal physical or mental strain, and all data, including recorded videos, were securely managed.

Experimental design and procedure

To evaluate the impact of adaptive question generation on the interviewees, each participant was interviewed. The participants were asked beforehand to describe key words that would be used as conversation starters. After the system first made introductory remarks to the participants, it began the dialogue with the question, "What has happened to (topic) recently?".

Measures

After the dialogue experiment, the participants were asked to answer a questionnaire and to perform a self-annotation of their own response utterances during the dialogue. In the self-annotation, annotations were given by the same questions as in Hazumi1911 for each exchange in which a question by the system was paired with an answer by the participant.

Dialogue breakdown annotation(preliminary experiment only): Based on the dialogue records of the preliminary experiment, annotations were assigned to whether the interview dialogue had broken down due to the system's inability to ask appropriate question utterances.

The following were annotated as "occurrence of dialogue breakdown" as annotation criteria.

- The speech that was not a question was made.
- The exact same question was repeated three or more times.

The reason for setting the number of times the same question is repeated at three times for the second reason for "dialogue breakdown" is that if it is two times in a row, the interviewee tends to guess that "robot looks want to hear more details " and continue the conversation well, and this is excluded from dialogue breakdown.

Self-disclosure scale:
The questions on the self-disclosure scale[58] present participants with 24 topics, and for each of the topics presented, participants rate on a 7-point scale how much they feel comfortable talking to the system. The questions presented are given four levels of self-disclosure (hobbies (Level 1), difficult experiences (Level 2), inconclusive shortcomings or weaknesses (Level 3), and negative personality or abilities (Level 4)), with the higher level requiring more in-depth self-disclosure. The self-disclosure score is calculated as an average value for each level.

Godspeed questionnaire:
Godspeed questionarie[65] evaluates five categories of human impressions of robots: anthropomorphism, animacy, likability, perceived intelligence, and perceived safety. Pairs of conflicting adjectives (e.g., "Artificial - Lifelike") are given in the questions, and participants rate which they are closer to in each pair. Scores are obtained as averages for each category.

Interviewer impression scale:
The Interviewer impression scale is a set of questions used in chapter 3 to evaluate the impression that the dialogue system gave to participants as an interviewer. Two questions (CQ1: Did you feel that the robot was interested in your answers in the interview? CQ2: Did you feel that the robot was asking you questions about topics you did not want to answer? This score is used to analyze the impact of the differences between the previous study and this study's system on impressions.

Post-annotation:
In the self-annotation, the examinees were asked to watch the video recording of the dialogue themselves, and for each dialogue exchange, three categories of questions (similar to Hazumi1911 (self-sentiment, topic-continue) and chapter 3 and (similar willingness annotations) were conducted. Questionnaire evaluation and self-annotation were conducted at the end of each interaction with each system.

Third-party dialogue evaluation:

We presented the third-party evaluator with the transcriptions of the dialogue content of the interview dialogues conducted in this experiment and the transcriptions of the dialogues conducted in the experiment in chapter 3, and asked them to score them. The annotators answered questions about fluency[62], naturalness as a response[75], consistency of context and relation[76], and godspeed[65] for each dialogue session.

The number of transcriptions to be evaluated was 30 for each of the old and new systems, for a total of 60, and it was concealed which entries were from which system. The transcriptions were replaced with generic nouns and pronouns to conceal which system they were from or the personal information of the participants. In addition, the order was randomized to reduce the impact of order.

Analysis

The purpose of this research is to improve the impression of the system on the interviewee and to increase the self-disclosure of the interviewee by adaptive questioning behavior based on the recognition of the motivation to speak. And to investigate the method of constructing a motivation to speak estimation model suitable for such a system.

The analysis in this experiment was conducted as follows in response to each research question.

Does dialogue breakdown cause willingness?

Before conducting a verification of the interview dialogue system incorporating adaptive question generation, we first verified through preliminary experiments that the system was functioning sufficiently. Based on the data obtained from this preliminary experiment, we evaluated the impact of the system not working properly. We observed the time series transition of the dialogue logs collected in the preliminary experiment and the dialogue breakdown annotations attached to them. We analyzed how the willingness assigned in the post-annotation changed before and after the timing of the dialogue breakdown annotation, as well as the trend of the occurrence of dialogue breakdowns.

Does LLM-based speech generation improve the quality of dialogue?:

We compared the Third-party dialogue evaluation scores assigned to the transcripts of the dialogue in this experiment and the transcripts of the dialogue in chapter3, respectively.

Does adaptive dialogue strategy encourage interviewees to self-disclosure?:

One of the most important hypotheses of this study is that adaptive dialogue strategies encourage self-disclosure by interviewees. To verify this hypothesis, we evaluated the differences between the two adaptive dialogue strategy systems and the random dialogue strategy system by conducting a Wilcoxon signed-rank test on the Self-disclosure scale collected after the dialogue.

Does adaptive dialogue strategy improve the impression given to the interviewee?:
Furthermore, it is also an important hypothesis that the impression given to the interviewee is improved through adaptive dialogue. We evaluated the differences between the two adaptive dialogue strategy systems and the random dialogue strategy system by conducting a Wilcoxon signed-rank test on the questionnaire evaluation scale (Godspeed questionnaire, Interviewer impression scale) collected after the dialogue.

How does the accuracy of attitude estimation affect the effectiveness of adaptive dialogue strategies?:
The third hypothesis is that the better the performance of the attitude estimation, the more effective the adaptive dialogue strategy will be. In order to evaluate the effect of attitude estimation accuracy on the effectiveness of adaptive dialogue strategies, we calculated the rate of agreement between the estimates made during the experiment and the self-annotations. We defined the rate of agreement between the attitude estimation results during the dialogue experiment and the results of the self-annotations made by the participants as the "correct rate", and conducted a correlation analysis between the correct rate and the results of the questionnaire evaluation.

How do differences in the target of user adaptation affect the effectiveness of adaptive dialogue?:
The fourth hypothesis is that there will be a difference in the effect of adaptive dialogue when the annotation of the attitude recognition to be adapted is conducted by the interviewee himself/herself or by a third party.)) by comparing the difference between the SS model and the TC model to evaluate the impact of the training data on the person annotations and the third-party annotations, respectively.

4.4 Result

4.4.1 Evaluation of multimodal attitude estimation models

Table 4.1 shows the results of cross-validation for the model SS and TC models. In-corpus shows the results of cross-validation within the Hazumi1911 corpus, and Out-corpus shows the estimation accuracy for the interview dia-

Table 4.1: Evaluation results for the combination of corpora and feature sets and the accuracy of attitude estimation

corpus	features	acc(SS)	var(SS)	acc(TC)	var(TC)
In-corpus	P+A	0.588	0.006	0.718	0.005
In-corpus	P+A+F+B	0.607	0.003	0.724	0.002
Out-corpus	P+A	0.538	0.060	0.607	0.084
Out-corpus	P+A+F+B	0.599	0.038	0.636	0.028

logue experiment corpus using the model trained on the Hazumi1911 corpus. acc is the estimation accuracy for each model, and var indicates the variance of the estimation accuracy. (SS) is the result of the SS model, which estimates self-sentiment, and (TC) is the result of the TC model, which estimates topic-continue.

The accuracy of the out-corpus was lower than that of the in-corpus, and the variance increased significantly. This result shows that the estimation accuracy decreases in an environment outside the corpus, and that there is a tendency for individual differences in accuracy to increase. Although the accuracy of TC was higher in the in-corpus, the accuracy of SS decreased by a smaller margin than TC, which decreased by nearly 10 points in the out-corpus.

In both in-corpus and out-corpus, and in both SS and TC, the more features used, the smaller the variance tended to be.

The variance for the P+A was 0.06 to 0.08 higher for the Out-corpus than for the In-corpus. On the other hand, for P+A+F+B, the increase in Out-corpus was 0.02 to 0.03 points higher than In-corpus, and the increase was smaller than for P+A. These results suggest that the greater the feature used, the smaller the individual differences in attitude estimation, and that it is also possible to suppress the increase in individual differences when operating outside the corpus.

4.4.2 Evaluation of self-disclosure through adaptive question generation and two types of user adaptation

Does dialogue breakdown cause willingness? :

We investigated the impact of adaptive question generation on the interviewees ' willingness. Figure 4.3 presents several examples of dialogue. In the graph, the horizontal axis represents the time elapsed since the start of the dialogue, while the vertical axis indicates the High/Low status of willingness. The estimated willingness results are marked with white circles, and the actual evaluations of willingness based on self-annotation are indicated with crosses. Points of dialogue breakdown are shown with triangles. The results of the graph revealed several trends. When the estimated willingness was initially Low, the actual willingness often also transitioned to Low. Furthermore, if the estimated willingness is continuously Low, it commonly results in the actual willingness eventually becoming Low. Even if the actual willingness later became High, a persistent Low estimate tended to lead to a quick decrease in willingness. Additionally, even when initial willingness was high, it could decrease due to dialogue breakdowns. Once a dialogue breakdown occurred, continuous breakdowns tended to ensue. This is thought to be due to the fact that the system's speech records are also included in the prompts.

From these results, we can see that maintaining the quality of system speech generation in adaptive dialogue systems and avoiding the occurrence of dialogue breakdowns is important for maintaining the interviewee's willingness. Once a breakdown in dialogue occurred, it continued to happen. This shows that it is important to take measures to prevent breakdowns in dialogue from occurring in the first place, and to replace them with alternative questions when they do occur.

Does LLM-based speech generation improve the quality of dialogue?：

The results of the comparison of the third-party dialogue evaluation are shown in Table 4.2. "LLM" is the improved system (LLM system : the system with question generation using LLM), and "Question list" is the baseline system(Question list system: the system selects questions from list constructed in chatper 3). The table shows the average of each question item, the p-value of the t-test result, and the difference in the average value.

The results showed significant differences in terms of the naturalness of the response sentences, consistency of context and relationship, fluency, and animacy, likability, and perceived intelligence, and the LLM system scored higher. The results suggest that LLM-based question generation improves the naturalness of speech and also achieves more contextually-matched dialogue, which gives a better impression.

On the other hand, there was no significant difference in the scores for anthropomorphism and perceived safety. This result shows that even when the responses are not natural or consistent with the context, there is no significant difference in the degree of anthropomorphism or safety perceived by a third party when using speech from a prepared list of questions. In

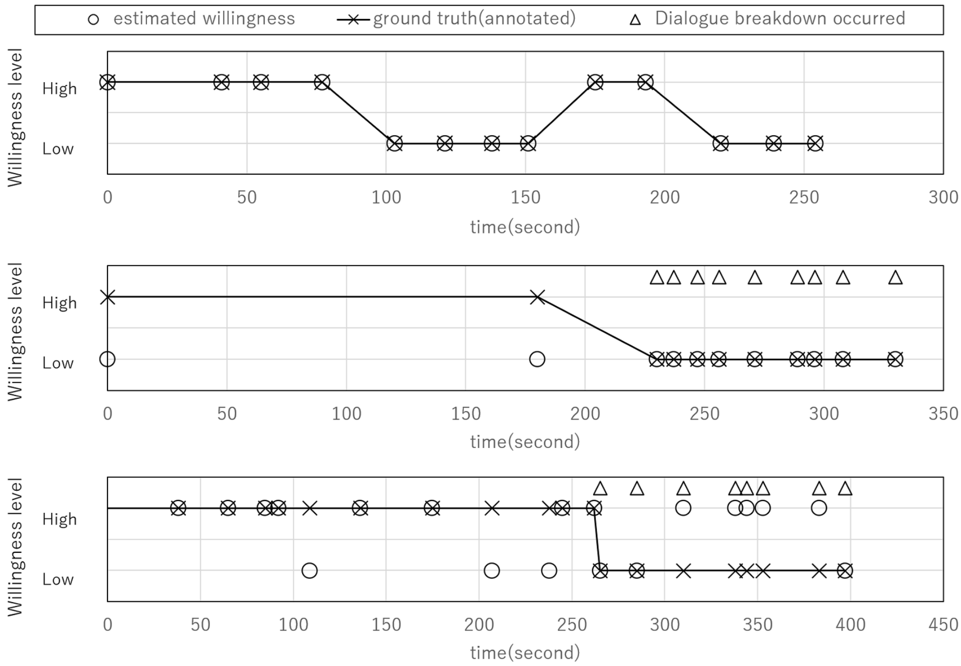

Figure 4.3: Timeline of willingness and Dialogue Breakdown

terms of perceived safety, since this evaluation was carried out by a third party, it was not possible to evaluate situations where the interviewee felt that they were being pryed into things they did not want to be asked about, and this is thought to be the reason why there was no difference in the scores.

Table 4.2: Dialogue quality as evaluated by a third party

		with LLM	Question list	p-value	difference in average values(LLM-list)
	naturalness as a response	3.37	2.67	0.002	0.70
	consistency of context and relations	3.67	3.22	0.041	0.45
	fluency	3.43	2.87	0.006	0.57
godspeed	Anthropomorphism	2.25	2.16	0.369	0.10
	Animacy	3.10	2.27	0.000	0.83
	Likability	3.14	2.41	0.000	0.72
	Perceived intelligence	3.03	2.34	0.000	0.69
	Perceived safety	2.76	2.79	0.470	-0.04

Does adaptive dialogue strategy improve the impression given to the interviewee?:

The means of responses to the Niwa scale and the Godspeed questionarie, as well as to the same questions as in the previous study, and the results of the cross-system tests for each response item are shown in Table 4.3. The response means for system(I)(Adaptation strategy in the Self-Sentiment model) are shown in column SS, for system(II)(Adaptation strategy in the Topic-continue model) in column TC, and for the Random strategy in column RND. As the results of the between-systems tests, the p-values obtained by the Wilkinson's signed rank test are shown in columns p(SS-RND), p(TC-RND), and p(TC-SS).

Results showed significant differences in some results between the adaptive and random strategy results. Regarding the depth of self-disclosure, significant differences were found between Level 2 and Level 3 for both p(SS-RND) and p(TC-RND).

Regarding the Godspeed questionarie assessment results, Significant differences were found for Animacy and Likeability between SS and RND, and for Perceived Intelligence between TC and RND. These results indicate that the adaptive dialogue strategy improved the interviewer's self-disclosure and impression on the interviewee (RQ1).

On the other hand, for Attitude of interest and Unpleasant question, significant differences were found only between SS and RND for Unpleasant question. No significant differences were found for Attitude of interest between any of the conditions in this study.

Table 4.3: Dialogue strategy and questionnaire results

		SS	TC	RND	p(SS-RND)	p(TC-RND)	p(SS-TC)
Self-disclosure scale	Self-disclosure Level 1	5.09	4.96	4.78	0.09	0.08	0.44
	Self-disclosure Level 2	3.81	3.77	3.27	**0.00**	**0.02**	0.49
	Self-disclosure Level 3	3.89	3.89	3.52	**0.03**	**0.05**	0.36
	Self-disclosure Level 4	3.53	3.56	3.24	0.19	0.09	0.36
Godspeed questionnaire	Anthropomorphism	2.89	2.88	2.63	**0.01**	0.06	0.42
	Animacy	3.35	3.33	2.99	**0.00**	**0.00**	0.20
	Likeability	3.88	3.83	3.68	**0.03**	0.09	0.11
	Perceived Intelligence	3.49	3.50	3.36	0.07	0.16	0.39
	Perceived Safety	2.64	2.70	2.70	0.19	0.47	0.26
Interviewer impression scale	Attitude of interest(small is better)	2.77	2.67	2.73	0.43	0.42	0.31
	Unpleasant question (large is better)	4.10	3.83	3.63	**0.02**	0.13	0.20

Table 4.4: Correlation coefficient between correct rate and questionnaire answers

	SS	TC
Percentage of utterances with high-willingness	0.303	0.379
Self-disclosure Level 1	0.010	0.180
Self-disclosure Level 2	0.072	0.045
Self-disclosure Level 3	0.325	0.191
Self-disclosure Level 4	0.151	0.183
Anthropomorphism	−0.021	0.031
Animacy	0.324	0.490
Likeability	0.366	0.337
Perceived Intelligence	0.188	0.406
Perceived Safety	0.003	−0.222
Attitude of interest(small is better)	0.193	−0.071
Unpleasant question (large is better)	0.249	0.137

How does the accuracy of attitude estimation affect the effectiveness of adaptive dialogue strategies? :

The percentage of agreement with the annotation results and the correlation coefficients with the agreement are shown in Table4.4. For the percentage of utterances with high-willingness, a slight correlation was found for both SS and TC. This result indicates that the proportion of motivated utterances increases as the accuracy of willingness recognition increases(RQ2).

Self-disclosure Level 3, Animacy, Likeability, and Perceived Intelligence were somewhat correlated with accuracy. These results suggest that improving the accuracy of the willingness recognition improves the effectiveness of adaptive interview dialogue.

Only for the “Unpleasant question (bigger is better),” a correlation was found in the SS model, but not in the TC model. This result suggests that the impression of being pursued for something one does not want to be asked can be avoided only when adaptive dialogue is conducted using a model trained with the self-annotated label.

How do differences in the target of user adaptation affect the effectiveness of adaptive dialogue? :

This study compared models using the “self-sentiment”and “topic-continue” labels from the Hazumi1911 data corpus for the purpose of assessing differences in adaptation between internal self-annotation and third-party annotation.

The results obtained in this study revealed some differences, particularly in terms of effects in adaptive strategies(RQ3). The SS model yielded significantly higher scores on more items of the godspeed rating scale; these differences were expressed despite the fact that SS and TC had generally equal values for the correct rate. One possible reason for this is that there were greater individual differences in effectiveness when the TC model was used, and fewer individual differences in effectiveness when the SS model was used.

On the other hand, when the correlation coefficients between the target and evaluation scores were compared, Self-disclosure Leve 3 showed correlation only in SS, while Perceived Intelligence showed correlation only in TC. This result suggests that the higher the accuracy of the self-annotation label, the deeper the self-disclosure, and the higher the accuracy of the third-party annotation, the more the interviewee gives the impression of approaching the interviewer as if he/she were a human being.

Using the same Hazumi data corpus, Katada et al.'s study [14] revealed that there are differences in useful features, with biometric features being more important for the self-annotated label estimation and visual features being more important for the third-party annotated label estimation.

In the present study, physiological signals were also highly important in the self-annotated model, while facial features analyzed from the captured videos were highly important in the third-party annotated model. Although this study used a different machine learning model from that of Kata et al.'s study, the trends of useful features were similar.

4.5 Discussion

4.5.1 User adpatation target and adaptive dialogue effects

This study compared models using the "self-sentiment" and "topic-continue" labels from the Hazumi1911 data corpus to assess differences in adaptation between internal self-annotation and third-party annotation.

The results obtained in this study revealed some differences, particularly in terms of effects on adaptive strategies (RQ3). The SS model yielded significantly higher scores on the godspeed questionnaire; these differences were observed although the SS and TC models generally had equal values for the correct rate. One possible reason for this finding is that there were greater individual differences in effectiveness when the TC model was used and fewer individual differences in effectiveness when the SS model was used. On the other hand, when the correlation coefficients between the target and evaluation scores were compared, self-disclosure Level 3 was correlated only with SS, whereas perceived intelligence was correlated only with TC. This result suggests that the greater the accuracy of self-annotation is, the greater the degree of self-disclosure will be; conversely, the greater the accuracy of third-party annotation is, the more likely the interviewee will be to approach the interviewer as if it was a human being.

4.5.2 Multimodal attitude estimation

Using the same Hazumi data corpus, Katada et al.'s study[14] revealed that there were differences in the importance of certain , with biometric features being more important for self-annotated label estimation and visual features being more important for third-party annotated label estimation. In this study, physiological signals were also highly important in the self-annotated model, whereas facial features analyzed from the captured videos were highly important in the third-party annotated model. Although this study used a different machine learning model from that of Kata et al.'s study, the trends of useful features were similar.

4.5.3 Self-disclosure

The results in Table 4.3 show that there was a significant difference between the adaptive strategy and the random strategy at Self-disclosure Levels 2 and 3, with the scores being higher in the case of the adaptive strategy. But, Self-disclosure was not promoted at all levels, and there was no significant difference at Levels 1 and 4.

One possible reason for this result is that the self-disclosure at Level 1 is about harmless topics such as hobbies, so the degree of self-disclosure did not change regardless of the dialogue strategy. There was also no significant difference in the case of Level 4, which is the deepest level of self-disclosure. In addition, the results in Table 4.4 show that even if the estimation accuracy is improved, there is no hope of improving the score for Level 4 self-disclosure. From this, it can be thought that the proposed method in this study is not sufficient for promoting the deepest level of self-disclosure, and some other method is needed.

4.6 Conclusion

In this chapter, as a further investigation of the adaptive interview dialogue robot system constructed and evaluated in chapter3, we investigated a multimodal attitude estimation model, explored the generation of question utterances, and evaluated the impact of self-disclosure through adaptive dialogue using an improved system.

The improved interview dialogue robot system uses an attitude estimation model trained using the Hazumi1911 dialogue corpus to decide whether to change the topic of each question or continue based on the estimated attitude, and automatically generates the next question using LLM.

First, as a further investigation of the multimodal attitude estimation model, we analyzed the changes in estimation accuracy and individual differences in estimation accuracy when the usage features were expanded and when the data was outside the training corpus. As a result, it was shown that the larger the usage features,the smaller the individual differences in attitude estimation could be made, and that the increase in individual differences when operating outside the corpus could also be suppressed.

Next, we implemented an improved system that incorporated a trained multimodal attitude estimation model and question generation using LLM. As a result of conducting dialogue experiments with 30 participants, the following findings were obtained.

When comparing the self-disclosure of interviewees in an interview dialogue robot system based on the evaluation scale by Niwa et al., it was found that the adaptive dialogue strategy, which performs adaptive topic follow-up based on multimodal attitude estimation, promotes self-disclosure in the second and third stages out of four stages. In addition, compared to random topic follow-up, the adaptive dialogue strategy improved the impression given to the interviewee.

There was a weak correlation between the estimation accuracy of multimodal attitude estimation and the scores for some items in self-disclosure and impression evaluation. In other words, it was found that improving the attitude recognition model increases the effectiveness of adaptive dialogue strategies.

In addition, it was found that the effect of adaptive dialogue caused some differences in impression evaluation depending on the attitude label (the person's own feelings or the attitude observed from the outside) that was the target of user adaptation.

Chapter 5

General Discussion

In this chapter, we will discuss the two research rounds that are a feature of this study, as well as the considerations that should be made by looking at the whole study, the limitations of this study, and future issues.

5.1 Comparison of the research rounds

This study began by building a system and conducting experiments. Based on the issues identified, we made improvements and conducted further experiments, and then made further improvements to the system in response to the issues that were identified in the experiments. For this reason, the overall study is divided into two major research rounds (chapter3 and chapter4). The experiments conducted in these two rounds differ in terms of the target of adaptation for inner-states and the way in which system utterances are created.

5.1.1 Differences in the inner-states to be adapted to

In the system of chapter3, we estimated willingness, and in chapter4, we estimated self-sentiment(SS) and topic-continue(TC).

In chapter4, it was shown that the proportion of self-disclosure and motivated speech by the interviewee increased significantly for both SS and TC. In that the proportion of motivated speech increased for all three, it can be said that the adaptive dialogue strategies targeted by this research are effective in all user adaptations. However, the impression evaluation items that correlated between SS and TC were different. This suggests that the impression and effect on the interviewee when user adaptation is performed for willingness may also differ.

The results of the analysis of the attitude estimation model, strategy and self-disclosure scores (section 4.4.2) showed that, although there was no significant difference, the Self-Sentiment estimation model obtained higher scores for a greater number of self-disclosure levels, and the correlation analysis results (section 4.4.2) of the attitude estimation accuracy and self-disclosure scores also showed that the SS model tended to have a higher correlation coefficient overall.

From these results, it can be thought that in order to encourage self-disclosure, it is more suitable to choose topics that match the sentiments that the person themselves feels, rather than the attitude that is observed from the outside. On the other hand, the estimation accuracy of the SS model tended to be lower than that of the TC model, which had the same conditions for feature values and corpus, so it is thought that in some cases, the estimation based on observed attitudes may be advantageous in terms of the ease of creating a highly accurate attitude estimation model. The question of which aspects of the user's internal state should be adapted to should be considered in light of the objectives of various tasks and systems.

5.1.2 Effects using LLM question generation

The main difference between R1 and R2 is that R1 selects questions from a pre-made list, while R2 automatically generates questions using LLM. In the system implemented in R1 (the question list system), all questions were placed on a tree structure graph based on topic relevance, and the system searched the tree using depth-first search and breadth-first search, switching between the two as needed to generate questions. A single common graph was prepared, and responses to various topics were handled using slotted filling. On the other hand, in the R2 system (LLM system), the next question is asked based on the latest five dialogue records, following prompts that include instructions for deepening or changing the topic.

The evaluation results for the differences in impressions before and after the introduction of LLM, obtained in section4.4.2, showed significant differences in the naturalness of the response sentences, consistency of the relationship with the context, fluency, and in the scores for Animacy, Likability, and Perceived intelligence, with the LLM system scoring higher. On the other hand, there were no significant differences in the scores for Anthropomorphism and Perceived safety. The results show that the LLM-based question generation improved the naturalness of the speech and also resulted in more contextually-matched dialogue, which gave a better impression.

On the other hand, there was no significant difference in the scores for anthropomorphism and perceived safety. The fact that there was no difference in anthropomorphism indicates that, when using a question list, even if the naturalness of the question utterances or their consistency with the context is low, there is no significant difference in the degree of anthropomorphism or safety perceived by a third party due to the naturalness of the utterances themselves. This is thought to be because, as pointed out in Inaba's discussion [64], the participants evaluated the highly crafted question list because they were evaluating a dialogue that lasted at most a few minutes.

With regard to perceived safety, since this evaluation was carried out by a third party, it was not possible to evaluate situations where the interviewee felt that they were being asked questions about things they did not want to be asked about, for example, and it may be said that this is why there was no difference in the scores.

Regarding the impression scale, the attitude of interest (smaller is better) score was significantly improved by the adaptive strategy in the question list system, but no such trend was seen in the LLM system. In addition, only the SS model showed a correlation for the unpleasant question (large is better), while no correlation was seen in the TC model. These results suggest that the impression of being pursued about something you don't want to be asked about can be avoided only when adaptive dialogue is conducted using a model trained with the person's own sentiments.

5.2 Random strategy as a baseline

We used a random strategy, which randomly continued or changed the topic regardless of the results of the internal state estimation, as a baseline system for comparison with the adaptive dialogue strategy. The experimental results show that the adaptive dialogue strategy increased the interviewees' willingness and promoted self-disclosure compared to the random strategy. There are other dialogue strategies that can be used as a baseline for comparison: the all-continue strategy(in which the topic is continued in all questions), and the all-change strategy(in which the topic is changed in all questions).

The all-continue strategy always delves deeper, so it will end up delving endlessly into only one topic. Since it is unclear which topic should be delved into for each interviewee, no questions will be asked about topics other than those that were covered in the first question. As a result, each interviewee will be polarized into two groups: those who are always highly motivated (to continue with topics that are with high-willingness) and those who are always less motivated (to continue with topics that are with less-willingness). As the topics that each interviewee was highly motivated about and the topics that they were less motivated about are different, if the interviewees are divided into a “ highly motivated group ” and a “ less motivated group ”, it will be difficult to make a fair comparison due to individual differences. The all-change strategy does not involve any in-depth probing. For this reason, questions that require self-disclosure are not asked, and it is difficult to evaluate whether self-disclosure actually took place. Also, if the topic is continually changed (especially in Round 1, as the top-level topics are limited), the questions will run out quickly, and it will not be possible to collect sufficient dialogue data.

For this reason, it is important to have a good mix of topics that continue and change, and the random strategy is appropriate as a baseline for adaptive dialogue strategies.

5.3 Impressions of the robot and the user

In this research, we used two types of humanoid robots to play the role of conversational partners with users in an interview dialogue system. In the first round, we used Pepper, made by Softbank, and in the second round, we used Sota, made by VStone. Pepper is a large robot that is about 120 cm tall, while Sota is a tabletop robot that is about 25 cm tall. The robots were chosen based on factors such as the experimental environment and the availability of the robots at the time of the experiment. Both robots were controlled to only generate speech and perform simple automatic gestures, and the robots themselves were made to behave as much as possible in the same way, but it is thought that this may have had some effect on the impression given to the test subjects.

According to a survey compiled by Mara et al.[68], Pepper tends to have a high score for anthropomorphism among humanoid robots, and it has been reported that NAO, a small robot like Sota, tends to be highly rated in terms of likability.

The results of the comparison of dialogue quality (section 4.4.2) showed a difference in dialogue quality between Round 1 and Round 2. As the evaluators in the third-party dialogue evaluation were only presented with the dialogue transcripts, there was no possibility of differences in the appearance of the robot influencing the evaluation, but on the other hand, it cannot be said that there was no impact on the dialogue behavior of the interviewees.

On the other hand, some of the experimental results suggest that the motivational enhancement and self-disclosure promotion of the interviewees through adaptive dialogue strategies are effective regardless of the robot. The significant difference in the self-disclosure score was a comparison between dialogues using the same robot, and a correlation with the correct response rate was also seen in some items. In addition, the percentage of motivated speech was higher in the case of adaptive dialogue strategies in both R1 and R2.

We proposed an adaptive interview dialogue strategy as a method that can be shared regardless of the appearance of the interview agent. The results of this paper showed that the proposed method had the effect of improving the willingness to speak and improving the impression of both Pepper and Sota, but it is necessary to investigate which dialogue agent is best suited for adaptive interview dialogue, such as virtual agents and text chat, and this is a future issue.

5.4 Utilization of language information

All of the features used to estimate the interviewees' inner states in this study were non-verbal information, and no verbal information was used in the multimodal attitude estimation.

There are several reasons for this: speech recognition is computationally intensive, so it is difficult to achieve both real-time performance and accuracy. This study implemented the system as a standalone system to ensure reproducibility of the experiment, so this study did not use a cloud service. And this study aimed for a method that could be applied regardless of language. On the other hand, language information is an important modality that is used in many related studies ([19, 16, 12, 14]), and it is also considered to be useful for estimating attitudes in interviews.

In recent years, speech recognition models that combine high accuracy and processing speed, such as whisper[77] used in the experiments in Round 2, have appeared, and it is predicted that even faster and more accurate speech recognition methods will become available in the future. The use of such speech recognition results for attitude estimation is also a future issue.

5.5 Adaptive interviewing techniques other than question selection

This study conducted user adaptation for question selection among various techniques used in interview dialogue. The system conducted user adaptation based only on the linguistic content of the next question to be spoken, and did not handle voice tone or gestures in this study. Various studies [78, 79] have shown that various system behaviors other than linguistic behaviors, such as gestures and nodding, improve user impressions. The findings of this study on linguistic user adaptation can be combined with the findings of these non-verbal methods of user adaptation. Further promotion of self-disclosure through this combination is a topic for future research.

Chapter 6

Geneal Conclusion

This research developed a real-time dialogue robot system that incorporates adaptive dialogue strategies based on online internal state estimation and evaluated the impact of these adaptive strategies on the overall dialogue. In this chapter, the conclusions from this research are summarized and the direction of future research is described.

First, an online willingness estimation model was trained and evaluated. These models use multimodal features through machine learning to estimate the interviewee's willingness. The models were trained to output the degree of motivation (high or low) for use in the question strategy of online adaptive dialogue strategies. Machine learning models were trained using random forests and LinearSVM as trainers with the postural features in the dialogues and the prosodic features of the speech as input, and the accuracy of the multimodal willingness recognition model was evaluated by cross-validation using the interview corpus. The results showed that the models correctly estimated the degree of willingness (high or low) with up to 72.8% accuracy. To address the issue of individual differences affecting estimation accuracy, especially when normalization data is unavailable (as in online scenarios), A pseudo-normalization method is proposed. This method effectively reduces the impact of individual differences on estimation accuracy and achieves more reliable real-time estimation.

Second, an interview dialogue robot system with an adaptive dialogue strategy that selects questions based on the estimated willingness using a trained willingness estimation model was constructed and evaluated. 27 interviewees were interviewed using two interview robot systems: (I) with an adaptive question selection module based on estimated speech willingness and (II) with a random question selection strategy. After the dialogues, impressions were evaluated using a questionnaire, as well as speech willingness annotations on the content of the dialogues by the interviewees themselves.

The results of experiment showed that the adaptive dialogue strategy gave the interviewees the impression that the robot was “more interested” in listening to them, and the adaptive questioning strategy significantly increased the percentage of utterances with high-willingness. This indicates that adaptive question selection using online willingness recognition was able to motivate the speaker even when the model could not be estimated with perfect accuracy.

Thirdly, the changes in accuracy due to the addition of features and the changes in individual differences in posture estimation accuracy were evaluated as improvements in the inner-state estimation used in the system. The learning and evaluation of a multimodal attitude estimation model trained on the Hazumi1911 dialogue corpus were carried out. The results showed that the addition of features improved the accuracy of attitude estimation and reduced the difference in estimation accuracy between individuals. In addition, to evaluate the impact on attitude estimation accuracy when a dialogue system is applied to the real world, the estimation accuracy between corpora in different sensing environments was evaluated. The estimation accuracy of the model trained on the Hazumi1911 dialogue corpus was evaluated on a newly collected interview dialogue corpus. The results showed that while the individual differences in estimation accuracy and precision worsened, the degree of this worsening varied depending on the label being estimated (whether it was a self-perception label or an attitude label assigned by a third party).

Fourth, question generation using LLM was proposed to enable interview dialogue based on adaptive dialogue strategy to be applied to arbitrary topics. In this study, the real-time generation of question utterances using a GPT model and its effectiveness were evaluated. In the proposed method, the real-time generation of question utterances is realized by dynamically replacing a part of the prompts. As a risk in real-time question generation, this study evaluated the effect on the willingness when a non-question utterance is generated, and confirmed that the interviewer's willingness decreases when a system utterance that is clearly not a question is generated and the flow of the question-and-answer session is disrupted. As a result of a dialogue experiment with 30 participants using an interview robot dialogue system with adaptive question generation, self-disclosure was promoted when adaptive question generation was used to continue/change the topic based on the estimated willingness, compared to the case where the topic was randomly continued/changed. In addition, a weak correlation was found between the accuracy of willingness estimation and impression evaluation scores, indicating that improving the accuracy of willingness estimation enhances the effectiveness of adaptive dialogue strategy.

This series of research results shows that a dialogue robot that can appropriately estimate the user's willingness and adaptively respond can promote user self-disclosure and provide a better interview experience, but based on the results to date, further studies on inner state estimation using multimodal features and intelligent However, based on the results to date, further studies on the estimation of internal states using multimodal features and intelligent dialogue processing based on such estimation remain as future research topics.

This study used a random forest as the online speech willingness recognition model, and this model was able to correctly estimate high or low of willingness with an accuracy of approximately 72%. Although this estimation accuracy is high compared to the chance level of the binomial classification task, the accuracy is low compared to study on inner-state estimation[14], such as sentiment estimation and engagement, and there is much room for improvement. The results in chapter 4 showed a weak correlation between the accuracy of attitude estimation and the effectiveness of adaptive dialogue strategy on dialogue, so the development of a method to estimate attitudes with higher accuracy is an important issue in improving the effectiveness of adaptive dialogue strategy.

This study proposed an adaptive dialogue strategy that determines whether to continue or change the topic based on the estimated internal state. This study also designed the estimation of the level of attitude as a binary classification task based on this. However, the willingness, self-sentiment, and topic-continue labels were each designed with five levels, and the fine differences were lost by binarization. The construction of an adaptive dialogue strategy that corresponds to the continuous strength and weakness of the internal state is a future issue.

In this study, the goal was to promote adaptive dialogue aimed to promote self-disclosure by encouraging deeper storytelling in interview dialogue. However, in interview dialogue, it is important not only to promote the subject's self-disclosure, but also to actively encourage the interviewer to talk about the information the interviewer wants, and to evaluate how much knowledge was elicited from the subject. It is necessary to construct a dialogue strategy that not only enlivens the dialogue but also leads the dialogue in accordance with the objectives of the system's tasks, as well as a method to evaluate the impact of the user's behavioral change in response to the dialogue task.

Finally, the entire dialogue system, including the adaptive dialogue strategy, requires optimization and comprehensive evaluation. In the system developed in this study, in the slowest case, there was a delay of almost 8 seconds between the end of user speech and the robot's next speech. In the verification in this study, no difference in response time was observed between the systems because all systems used the same software module to generate utterances, but the effect of the response time of the dialogue system and the effect of the dialogue strategy, as well as the effect of the dialogue strategy depending on whether the dialogue interface is a robot or a text chat, etc., were analyzed. However, there is a need for further investigation and understanding of user adaptation to the overall behavior and performance of various dialogue systems, including analysis of the effects of dialogue strategy depending on the form of the dialogue interface (e.g., robot, text chat, etc.).

Thus, the advancement of multimodal inner state estimation and adaptive dialogue strategy, as well as a deeper understanding and exploration of intelligent dialogue processing, will lead to the realization of adaptive dialogue systems and emotionally intelligent agents.

Bibliography

[1] Townsend G., LaPallo B.K., Boulay C.B., Krusienski D.J., Frye G.E., Hauser C.K., Schwartz N.E., Vaughan T.M., Wolpaw J.R., and Sellers E.W. "A novel P300-based brain–computer interface stimulus presentation paradigm: Moving beyond rows and columns". *Clinical Neurophysiology* 121.7 (July 2010), pp. 1109–1120. ISSN: 1388-2457. DOI: 10.1016/j.clinph.2010.01.030. URL: https://cir.nii.ac.jp/crid/1361418520651649024.

[2] Boštjan Šumak, Matic Špindler, Mojca Debeljak, Marjan Heričko, and Maja Pušnik. "An empirical evaluation of a hands-free computer interaction for users with motor disabilities". *J. of Biomedical Informatics* 96.C (Aug. 2019). ISSN: 1532-0464. DOI: 10.1016/j.jbi.2019.103249. URL: https://doi.org/10.1016/j.jbi.2019.103249.

[3] Matthew B Hoy. "Alexa, Siri, Cortana, and more: an introduction to voice assistants". *Medical reference services quarterly* 37.1 (2018), pp. 81–88.

[4] Michael F McTear. "Spoken dialogue technology: enabling the conversational user interface". *ACM Computing Surveys (CSUR)* 34.1 (2002), pp. 90–169.

[5] Wolfgang Wahlster and Alfred Kobsa. "User models in dialog systems". *User models in dialog systems* (1989), pp. 4–34.

[6] Hongshen Chen, Xiaorui Liu, Dawei Yin, and Jiliang Tang. "A Survey on Dialogue Systems: Recent Advances and New Frontiers". *SIGKDD Explor. Newsl.* 19.2 (Nov. 2017), pp. 25–35. ISSN: 1931-0145. DOI: 10.1145/3166054.3166058. URL: https://doi.org/10.1145/3166054.3166058.

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

[8] Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. "Improving language understanding by generative pre-training" (2018).

[9] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. "Language models are few-shot learners". *Proceedings of the 34th International Conference on Neural Information Processing Systems*. NIPS '20. Vancouver, BC, Canada: Curran Associates Inc., 2020. ISBN: 9781713829546.

[10] Inyoung Cheong, Aylin Caliskan, and Tadayoshi Kohno. "Envisioning Legal Mitigations for Intentional and Unintentional Harms Associated with Large Language Models" ().

[11] Hannah R Lawrence, Renee A Schneider, Susan B Rubin, Maja J Matarić, Daniel J McDuff, and Megan Jones Bell. "The Opportunities and Risks of Large Language Models in Mental Health". *JMIR Ment Health* 11 (July 2024), e59479. ISSN: 2368-7959. DOI: `10.2196/59479`. URL: `https://doi.org/10.2196/59479`.

[12] Yuki Hirano, Shogo Okada, Haruto Nishimoto, and Kazunori Komatani. "Multitask Prediction of Exchange-Level Annotations for Multimodal Dialogue Systems". *Proc. International Conference on Multimodal Interaction (ICMI)*. 2019, pp. 85–94.

[13] Yuya Chiba, Takashi Nose, and Akinori 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* 2018-February.December (2018), pp. 428–431. DOI: `10.1109/APSIPA.2017.8282069`.

[14] Shun Katada, Shogo Okada, and Kazunori Komatani. "Effects of Physiological Signals in Different Types of Multimodal Sentiment Estimation". *IEEE Transactions on Affective Computing* (2022). DOI: `10.1109/TAFFC.2022.3155604`.

[15] Alessandro Vinciarelli. “Introduction: Social Signal Processing”. *Social Signal Processing*. Ed. by Judee K. Burgoon, Nadia Magnenat-Thalmann, Maja Pantic, and AlessandroEditors Vinciarelli. Cambridge University Press, 2017, pp. 1–8.

[16] Mohammed Ehsan Hoque, Matthieu Courgeon, Jean-Claude Martin, Bilge Mutlu, and Rosalind W Picard. “Mach: My automated conversation coach”. *Proc. International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp)*. ACM. 2013, pp. 697–706.

[17] Hiroki Tanaka, Hideki Negoro, Hidemi Iwasaka, and Satoshi Nakamura. “Embodied conversational agents for multimodal automated social skills training in people with autism spectrum disorders”. *PloS one* 12.8 (2017), e0182151.

[18] Andrius Dzedzickis, Artūras Kaklauskas, and Vytautas Bucinskas. “Human Emotion Recognition: Review of Sensors and Methods”. *Sensors* 20.3 (2020). ISSN: 1424-8220. DOI: 10.3390/s20030592. URL: https://www.mdpi.com/1424-8220/20/3/592.

[19] David DeVault, Ron Artstein, Grace Benn, Teresa Dey, Ed Fast, Alesia Gainer, Kallirroi Georgila, Jon Gratch, Arno Hartholt, Margaux Lhommet, Gale Lucas, Stacy Marsella, Fabrizio Morbini, Angela Nazarian, Stefan Scherer, Giota Stratou, Apar Suri, David Traum, Rachel Wood, Yuyu Xu, Albert Rizzo, and Louis-Philippe Morency. “SimSensei Kiosk: A Virtual Human Interviewer for Healthcare Decision Support”. *Proc. International Conference on Autonomous Agents and Multi-agent Systems (AAMAS)*. International Foundation for Autonomous Agents and Multiagent Systems, 2014, pp. 1061–1068.

[20] Masaharu Naito, Daniel J Rea, and Takayuki Kanda. “Hey robot, tell it to me straight: how different service strategies affect human and robot service outcomes”. *International Journal of Social Robotics* 15.6 (2023), pp. 969–982.

[21] Ben Emans. *Interviewing: Theory, techniques and training*. London: Routledge, 2016.

[22] Andrea Fontana and James H Frey. “The interview”. *The Sage handbook of qualitative research* 3.1 (2005), pp. 695–727.

[23] Frank Schmidt and J.E. Hunter. “The Validity and Utility of Selection Methods in Personnel Psychology: Practical and Theoretical Implications of 85 Years of Research Findings”. *Psychological Bulletin* 124 (Sept. 1998), pp. 262–274. DOI: 10.1037//0033-2909.124.2.262.

[24] Joseph Weizenbaum. "ELIZA—a computer program for the study of natural language communication between man and machine". *Commun. ACM* 9.1 (Jan. 1966), pp. 36–45. ISSN: 0001-0782. DOI: 10.1145/365153.365168. URL: https://doi.org/10.1145/365153.365168.

[25] Luísa Coheur. "From Eliza to Siri and beyond". *Information Processing and Management of Uncertainty in Knowledge-Based Systems: 18th International Conference, IPMU 2020, Lisbon, Portugal, June 15–19, 2020, Proceedings, Part I 18*. Springer. 2020, pp. 29–41.

[26] Dimitris Spiliotopoulos, Ion Androutsopoulos, and Constantine D Spyropoulos. "Human-robot interaction based on spoken natural language dialogue". *Proceedings of the European workshop on service and humanoid robots*. 2001, pp. 25–27.

[27] Shohei Tanaka, Koichiro Yoshino, Katsuhito Sudoh, and Satoshi Nakamura. "ARTA: Collection and Classification of Ambiguous Requests and Thoughtful Actions". *Proceedings of the 22nd Annual Meeting of the Special Interest Group on Discourse and Dialogue* (Jan. 2021). DOI: 10.18653/v1/2021.sigdial-1.9. URL: https://cir.nii.ac.jp/crid/1871146592926968064.

[28] Rahul Goel, Chris Hidey, Hafeezul Rahman Mohammad, Pavankumar Reddy Muddireddy, and Fei Liu. "LLM-based Task-oriented Dialog System with Few-shot Retrieval Augmentation" (2023).

[29] Angus Addlesee, Neeraj Cherakara, Nivan Nelson, Daniel Hernández García, Nancie Gunson, Weronika Sieińska, Marta Romeo, Christian Dondrup, and Oliver Lemon. "A Multi-party Conversational Social Robot Using LLMs". *Companion of the 2024 ACM/IEEE International Conference on Human-Robot Interaction*. HRI '24. Boulder, CO, USA: Association for Computing Machinery, 2024, pp. 1273–1275. ISBN: 9798400703232. DOI: 10.1145/3610978.3641112. URL: https://doi.org/10.1145/3610978.3641112.

[30] Alexei A. Birkun and Adhish Gautam. "Large Language Model (LLM)-Powered Chatbots Fail to Generate Guideline-Consistent Content on Resuscitation and May Provide Potentially Harmful Advice". *Prehospital and Disaster Medicine* 38.6 (2023), pp. 757–763. DOI: 10.1017/S1049023X23006568.

[31] A MEHRABIAN. "Communication without words". *Psychol. Today* 2.4 (1968), pp. 53–56.

[32] Sebastian Schneider and Franz Kummert. “Comparing the effects of social robots and virtual agents on exercising motivation”. *Social Robotics: 10th International Conference, ICSR 2018, Qingdao, China, November 28-30, 2018, Proceedings 10*. Springer. 2018, pp. 451–461.

[33] Jamy Li. “The benefit of being physically present: A survey of experimental works comparing copresent robots, telepresent robots and virtual agents”. *International Journal of Human-Computer Studies* 77 (2015), pp. 23–37. ISSN: 1071-5819. DOI: https://doi.org/10.1016/j.ijhcs.2015.01.001. URL: https://www.sciencedirect.com/science/article/pii/S107158191500004X.

[34] Koji Inoue, Divesh Lala, Kenta Yamamoto, Shizuka Nakamura, Katsuya Takanashi, and Tatsuya Kawahara. “An Attentive Listening System with Android ERICA: Comparison of Autonomous and WOZ Interactions”. *Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue*. Ed. by Olivier Pietquin, Smaranda Muresan, Vivian Chen, Casey Kennington, David Vandyke, Nina Dethlefs, Koji Inoue, Erik Ekstedt, and Stefan Ultes. 1st virtual meeting: Association for Computational Linguistics, July 2020, pp. 118–127. DOI: 10.18653/v1/2020.sigdial-1.15. URL: https://aclanthology.org/2020.sigdial-1.15.

[35] Takato Yamazaki, Katsumasa Yoshikawa, Toshiki Kawamoto, Tomoya Mizumoto, Masaya Ohagi, and Toshinori Sato. “Building a hospitable and reliable dialogue system for android robots: a scenario-based approach with large language models”. *Advanced Robotics* 37.21 (2023), pp. 1364–1381. DOI: 10.1080/01691864.2023.2244554.

[36] Chloé Clavel, Matthieu Labeau, and Justine Cassell. “Socio-conversational systems: Three challenges at the crossroads of fields”. *Frontiers in Robotics and AI* 9 (2022). ISSN: 2296-9144. DOI: 10.3389/frobt.2022.937825. URL: https://www.frontiersin.org/journals/robotics-and-ai/articles/10.3389/frobt.2022.937825.

[37] Alessandro Vinciarelli, Maja Pantic, and Hervé Bourlard. “Social signal processing: Survey of an emerging domain”. *Image and vision computing* 27.12 (2009), pp. 1743–1759.

[38] Nadine Glas and Catherine Pelachaud. “Definitions of engagement in human-agent interaction”. *Proc. International Workshop on Engagment in Human Computer Interaction (ENHANCE)*. 2015, pp. 944–949.

[39] Koji Inoue, Divesh Lala, Katsuya Takanashi, and Tatsuya Kawahara. "Latent character model for engagement recognition based on multimodal behaviors". *Proc. International Workshop on Spoken Dialogue Systems (IWSDS).* 2018.

[40] Yukiko I. Nakano and Ryo Ishii. "Estimating User's Engagement from Eye-gaze Behaviors in Human-agent Conversations". *Proc. International Conference on Intelligent User Interfaces (IUI).* ACM, 2010, pp. 139–148.

[41] Dan Bohus and Eric Horvitz. "Learning to Predict Engagement with a Spoken Dialog System in Open-world Settings". *Proc. Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL).* Association for Computational Linguistics, 2009, pp. 244–252.

[42] Candace Sidner, Cory Kidd, Christopher Lee, and Neal Lesh. "Where to look: A study of human-robot engagement". *Proc. International Conference on Intelligent User Interfaces (IUI).* Jan. 2004, pp. 78–84. DOI: `10.1145/964442.964458`.

[43] Catharine Oertel, Ginevra Castellano, Mohamed Chetouani, Jauwairia Nasir, Mohammad Obaid, Catherine Pelachaud, and Christopher Peters. "Engagement in Human-Agent Interaction: An Overview". *Frontiers in Robotics and AI* 7 (Aug. 2020). ISSN: 22969144. DOI: `10.3389/frobt.2020.00092`.

[44] Kazunori Komatani and Shogo Okada. "Multimodal Human-Agent Dialogue Corpus with Annotations at Utterance and Dialogue Levels". *2021 9th International Conference on Affective Computing and Intelligent Interaction (ACII).* 2021, pp. 1–8. DOI: `10.1109/ACII52823.2021.9597447`.

[45] Hiroki Tanaka, Hideki Negoro, Hidemi Iwasaka, and Satoshi Nakamura. "Listening Skills Assessment Through Computer Agents". *Proc. International Conference on Multimodal Interaction (ICMI).* ACM, 2018, pp. 492–496. ISBN: 978-1-4503-5692-3.

[46] Hiroki Tanaka, Hiroyoshi Adachi, Norimichi Ukita, Manabu Ikeda, Hiroaki Kazui, Takashi Kudo, and Satoshi Nakamura. "Detecting Dementia Through Interactive Computer Avatars". *IEEE journal of translational engineering in health and medicine* 5 (2017), pp. 1–11.

[47] Masahiro Araki, Sayaka Tomimasu, Mikio Nakano, Kazunori Komatani, Shogo Okada, Shinya Fujie, and Hiroaki Sugiyama. "Collection of Multimodal Dialog Data and Analysis of the Result of Annotation of Users' Interest Level". *Proc. International Conference*

on Language Resources and Evaluation (LREC). European Language Resources Association (ELRA), 2018.

[48] Sayaka Tomimasu and Masahiro Araki. "Assessment of Users' Interests in Multimodal Dialog Based on Exchange Unit". *Proc. International Workshop on Multimodal Analyses Enabling Artificial Agents in Human-Machine Interaction*. ACM, 2016, pp. 33–37.

[49] L. Batrinca, N. Mana, B. Lepri, N. Sebe, and F. Pianesi. "Multimodal Personality Recognition in Collaborative Goal-Oriented Tasks". *IEEE Transactions on Multimedia* 18.4 (2016), pp. 659–673.

[50] Klaus Weber, Hannes Ritschel, Ilhan Aslan, Florian Lingenfelser, and Elisabeth André. "How to Shape the Humor of a Robot - Social Behavior Adaptation Based on Reinforcement Learning". *Proc. International Conference on Multimodal Interaction (ICMI)*. ACM, 2018, pp. 154–162.

[51] Setareh Nasihati Gilani, David Traum, Arcangelo Merla, Eugenia Hee, Zoey Walker, Barbara Manini, Grady Gallagher, and Laura-Ann Petitto. "Multimodal Dialogue Management for Multiparty Interaction with Infants". *Proc. International Conference on Multimodal Interaction (ICMI)*. ACM, 2018, pp. 5–13.

[52] Naoko Saito, Shogo Okada, Katsumi Nitta, Yukiko I. Nakano, and Yuki Hayashi. "Estimating user's attitude in multimodal conversational system for elderly people with dementi". *AAAI Spring Symposium - Technical Report* SS-15-07 (2015), pp. 100–103.

[53] Giota Stratou and Louis-Philippe Morency. "MultiSense—Context-Aware Nonverbal Behavior Analysis Framework: A Psychological Distress Use Case". *IEEE Transactions on Affective Computing* 8.2 (2017), pp. 190–203. ISSN: 1949-3045.

[54] Sooyeon Jeong, Laura Aymerich-Franch, Kika Arias, Sharifa Alghowinem, Agata Lapedriza, Rosalind Picard, Hae Won Park, and Cynthia Breazeal. "Deploying a robotic positive psychology coach to improve college students ' psychological well-being". *User Modeling and User-Adapted Interaction* 33.2 (2023), pp. 571–615.

[55] Irwin Altman. "Social penetration: The development of interpersonal relationships". *Rinehart, & Winston* (1973).

[56] Mohammad Soleymani, Kalin Stefanov, Sin Hwa Kang, Jan Ondras, and Jonathan Gratch. "Multimodal analysis and estimation of intimate self-disclosure". *ICMI 2019 - Proceedings of the 2019 International Conference on Multimodal Interaction* (2019), pp. 59–68. DOI: 10.1145/3340555.3353737.

[57] Seiya Mitsuno, Midori Ban, Hiroshi Ishiguro, and Yuichiro Yoshikawa. "Deepening Conversations Over Time: A Chatbot with a Topic Depth Estimation Model for Gradually Engaging in Deeper Chats". *2024 33rd IEEE International Conference on Robot and Human Interactive Communication (ROMAN)*. 2024, pp. 1354–1361. DOI: 10.1109/RO-MAN60168.2024.10731430.

[58] Sora Niwa and Shun'ichi MARUNO. "Development of a Scale to Assess the Depth of Self-disclosure." *Japanese Journal of Personality* 18.3 (2010).

[59] Takahiro Kobori, Mikio Nakano, and Tomoaki 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)* September (2016), pp. 370–380. DOI: 10.18653/v1/w16-3646.

[60] Takuya Ishihara, Fuminori Nagasawa, Katsumi Nitta, and Shogo Okada. "Estimating interviewee's willingness in multimodal human robot interview interaction". *Proceedings of the 20th International Conference on Multimodal Interaction, ICMI 2018* (2018). DOI: 10.1145/3281151.3281153.

[61] Ashwin Ram, Rohit Prasad, Chandra Khatri, Anu Venkatesh, Raefer Gabriel, Qing Liu, Jeff Nunn, Behnam Hedayatnia, Ming Cheng, Ashish Nagar, Eric King, Kate Bland, Amanda Wartick, Yi Pan, Han Song, Sk Jayadevan, Gene Hwang, and Art Pettigrue. "Conversational AI: The Science Behind the Alexa Prize". *ArXiv* abs/1801.03604 (2018). URL: https://api.semanticscholar.org/CorpusID:8706813.

[62] Emily Dinan, Varvara Logacheva, Valentin Malykh, Alexander Miller, Kurt Shuster, Jack Urbanek, Douwe Kiela, Arthur Szlam, Iulian Serban, Ryan Lowe, Shrimai Prabhumoye, Alan W. Black, Alexander Rudnicky, Jason Williams, Joelle Pineau, Mikhail Burtsev, and Jason Weston. "The Second Conversational Intelligence Challenge (ConvAI2)". *The NeurIPS '18 Competition*. Ed. by Sergio Escalera and Ralf Herbrich. Cham: Springer International Publishing, 2020, pp. 187–208. ISBN: 978-3-030-29135-8.

[63] Ryuichiro HIGASHINAKA. “Evaluation of chat-oriented dialogue systems and its problems”. *JSAI Technical Report, SIG-FPAI* 101 (2016), p. 04. DOI: 10.11517/jsaifpai.101.0_04.

[64] Michimasa INABA. “How should We Evaluate Chat-oriented Dialogue Systems?” *JSAI Technical Report, SIG-SLUD* 87 (2019), p. 11. DOI: 10.11517/jsaislud.87.0_11.

[65] Christoph Bartneck, Dana Kulić, Elizabeth Croft, and Susana Zoghbi. “Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots”. *International journal of social robotics* 1 (2009), pp. 71–81.

[66] Astrid Weiss and Christoph Bartneck. “Meta analysis of the usage of the Godspeed Questionnaire Series”. *2015 24th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*. 2015, pp. 381–388. DOI: 10.1109/ROMAN.2015.7333568.

[67] Thunberg Sofia. “Investigating the Social Influence of Different Humanoid Robots” (2017), p. 28.

[68] Martina Mara, Markus Appel, and Timo Gnambs. “Human-Like Robots and the Uncanny Valley: A Meta-Analysis of User Responses Based on the Godspeed ScalesHuman-Like Robots and the Uncanny Valley: A Meta-Analysis of User Responses Based on the Godspeed Scales”. 1 (2022), pp. 33–46. ISSN: 2190-8370. URL: 10.1027/2151-2604/a000486.

[69] Koji Inoue, Kohei Hara, Divesh Lala, Kenta Yamamoto, Shizuka Nakamura, Katsuya Takanashi, and Tatsuya Kawahara. “Job Interviewer Android with Elaborate Follow-up Question Generation”. *Proceedings of the 2020 International Conference on Multimodal Interaction*. ICMI ’20. Virtual Event, Netherlands: Association for Computing Machinery, 2020, pp. 324–332. ISBN: 9781450375818. DOI: 10.1145/3382507.3418839. URL: https://doi.org/10.1145/3382507.3418839.

[70] SoftBank Robotics. *SOFTBANK ROBOTICS DOCUMENTATION NAOqi SDK*. "http : / / doc . aldebaran . com / 2 - 5/index_dev_guide.html".

[71] Dexter C Kozen. “Depth-first and breadth-first search”. *The design and analysis of algorithms*. Springer, 1992, pp. 19–24.

[72] Florian Eyben, Martin Wöllmer, and Björn Schuller. "Opensmile: The Munich Versatile and Fast Open-Source Audio Feature Extractor". *Proceedings of the 18th ACM International Conference on Multimedia*. MM '10. Firenze, Italy: Association for Computing Machinery, 2010, pp. 1459–1462. ISBN: 9781605589336. DOI: 10.1145/1873951.1874246. URL: https://doi.org/10.1145/1873951.1874246.

[73] Corinna Cortes and Vladimir Vapnik. "Support-Vector Networks". *Machine Learning*. 1995, pp. 273–297.

[74] Naoaki,Okazaki Sakae,Mizuki Hiroki,Iida Mengsay,Loem Shota,Hirai Kakeru,Hattori Masanari,Ohi Rio,Yokota Kazuki,Fujii Taishi,Nakamura. *Swallow-LLM*. Accessed: 2024-03-04. 2023.

[75] Hiroaki SUGIYAMA, Toyomi MEGURO, and Ryuichiro HIGASHINAKA. "Multi-aspect Evaluation for Utterances in Chat Dialogues". *JSAI Technical Report, SIG-SLUD* 72 (2014), p. 06. DOI: 10.11517/jsaislud.72.0_06.

[76] Tong Niu and Mohit Bansal. "Polite Dialogue Generation Without Parallel Data". *Transactions of the Association for Computational Linguistics* 6 (2018). Ed. by Lillian Lee, Mark Johnson, Kristina Toutanova, and Brian Roark, pp. 373–389. DOI: 10.1162/tacl_a_00027. URL: https://aclanthology.org/Q18-1027.

[77] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. "Robust speech recognition via large-scale weak supervision". *International conference on machine learning*. PMLR. 2023, pp. 28492–28518.

[78] Sarah Hanna Fischer. "Design and evaluation of non-verbal cues for the Robot Pepper". PhD thesis. Wien, 2021.

[79] Sin-Hwa Kang, Jonathan Gratch, Candy Sidner, Ron Artstein, Lixing Huang, and Louis-Philippe Morency. "Towards building a virtual counselor: modeling nonverbal behavior during intimate self-disclosure". *Proceedings of the 11th International Conference on Autonomous Agents and Multiagent Systems - Volume 1*. AAMAS '12. Valencia, Spain: International Foundation for Autonomous Agents and Multiagent Systems, 2012, pp. 63–70. ISBN: 0981738117.

Appendix

Sample of dialogue log from Round 2

Tables 6.1 and 6.2 show the transcriptions of the dialogue logs from the dialogue experiments conducted in Research Round 2 4. The "USER" speaker is the interviewee, and the "SYSTEM" is the interview robot system.

Table 6.1: Dialogue sample with LLM question generation

speaker	text
SYSTEM	What have you been up to recently?
USER	I'm going to Shikoku at the end of this month. Do you know Shikoku?
SYSTEM	Are you going on a trip to Shikoku? What kind of place is Shikoku?
USER	Well, I'm going to Kagawa Prefecture, which is in Kagawa Prefecture. I think it's warmer there than here, but I'm not sure. I have an image that the food is delicious, but...
SYSTEM	You have an image that the food in Kagawa Prefecture is delicious. What kind of food in Shikoku are you interested in?
USER	Well, when you go to Kagawa, you eat Sanuki Udon, right? It's something like chicken, but not yakitori. I've heard that there's a place famous for grilled chicken, so I'm thinking of going there.
SYSTEM	Do you want to eat Sanuki Udon in Kagawa Prefecture?
USER	I want to eat it. I'm going to stay at Hirahonsen, which is on the approach to the main gate of Konpira-san. I'm really looking forward to it. Also, do you know about the big joke and the small joke? I'd like to go there too, if I have time.

Table 6.2: A dialogue log showing a dialog breakdown due to a failed question. It shows the user looking confused when the same question is repeated three times.

speaker	text
	(USER taked that she collect antiques and other collectibles as a hobby.)
SYSTEM	Why do you like it?
USER	I like it because I feel the historical value, or the weight of history.
SYSTEM	Why do you like it?
USER	Well, yes. As I said before, I like it because I feel the weight of history and the sense of that era, and also, um, what was I going to say? Also, there are technologies that can't be made now, and things that have disappeared. They're called lost technologies, but I feel a sense of value in things that have been made with old technologies and have survived for a very long time.
SYSTEM	Why do you like it?
USER	That's true. Well... when you ask me why three times... when you ask me why... I'm not sure... but... well...

Publication List

International Journal (peer reviewed)

Nagasawa Fuminori, Okada Shogo, Ishihara Takuya, and Nitta,Katsumi. Adaptive Interview Strategy Based on Interviewees Speaking Willingness Recognition for Interview Robots. IEEE Transactions on Affective Computing. 2023.
(chapter 3)

International Conference (peer reviewed)

Nagasawa, Fuminori, and Shogo Okada. Investigating Effects of Multimodal Topic-continuance Recognition on Human-Robot Interviewing Interaction. Companion of the 2024 ACM/IEEE International Conference on Human-Robot Interaction. 2024.
(section4.4.2)

Domestic Conference

Nagasawa, Fuminori, and Shogo Okada. Interview robot dialogue system with adaptive dialogue strategy based on estimation of willingness. In Proceedings of the 100th Meeting of the Japanese Society for Artificial Intelligence, SIG on Language and Speech Understanding and Dialogue Processing (2024/02) (pp. 204-209).

Nagasawa, Fuminori, and Shogo Okada. Adaptive Dialogue Strategy Based on Speech Motivation Recognition in an Interview Robot Dialogue System. In Proceedings of the 38th Annual Conference of the Japanese Society for Artificial Intelligence (pp. 3R1OS13b02-3R1OS13b02). 2024.

(Preparation)

Nagasawa, Fuminori, and Shogo Okada. Can adaptive interviewer robot based on social signals make a better impression on interviewees and encourage self-disclosure?.
(chapter 4 except section 4.4.2)