



Scenario-based dialogue system based on pause detection toward daily health monitoring

Journal of Rehabilitation and Assistive Technologies Engineering
Volume 9: 1–23
© The Author(s) 2022
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/20556683221133367
journals.sagepub.com/home/jrt


Kazumi Kumagai^{1,2} , Seiki Tokunaga¹, Norihisa P Miyake¹, Kazuhiro Tamura¹, Ikuro Mizuuchi² and Mihoko Otake-Matsuura¹

Abstract

Introduction: We have conducted research on building a robot dialogue system to support the independent living of older adults. In order to provide appropriate support for them, it is necessary to obtain as much information, particularly related to their health condition, as possible. As the first step, we have examined a method to allow dialogue to continue for longer periods.

Methods: A scenario-based dialogue system utilizing pause detection for turn-taking was built. The practicality of adjusting the system based on the dialogue rhythm of each individual was studied. The system was evaluated through user studies with a total of 20 users, 10 of whom were older adults.

Results: The system detected pauses in the user's speech using the sound level of their voice, and predicted the duration and number of pauses based on past dialogue data. Thus, the system initiated the robot's voice-call after the user's predicted speech.

Conclusions: Multiple turns of dialogue between robot and older adults are found possible under the system, despite several overlaps of robot's and users' speech observed. The users responded to the robot, including the questions related to health conditions. The feasibility of a scenario-based dialogue system was suggested; however, improvements are required.

Keywords

robotics for ambient assisted living, human robot interaction, scenario-based dialogue, turn-taking

Introduction

In recent years, the advancement of a super-aged society has been accelerating at a rapid pace. Along with this, the shortage of care workers and the shortage of care facilities are becoming social problems, and thus the care support equipment and systems using information and communication technology (ICT), including robotic technology, are being intensively developed.^{1,2} The number of older adults living at home is increasing remarkably due to the shortage of care facilities, and the proportion of older adults living alone is increasing accordingly. Therefore, ensuring the quality of

their daily living and finding ways to manage their physical and mental health are becoming urgent social issues.³

¹Center for Advanced Intelligence Project, RIKEN, Nihonbashi 1-chome Mitsui Building, 15th floor, 1-4-1 Nihonbashi, Chuo-ku, Tokyo 103-0027, Japan

²Tokyo University of Agriculture and Technology, 2-24-16 Naka-cho, Koganei-shi, Tokyo 184-8588, Japan

Corresponding author:

Kazumi Kumagai, RIKEN, 2-1 Hirosawa, Wako 351-0198, Japan.
Email: kazumi.kumagai@riken.jp



Creative Commons Non Commercial CC BY-NC: This article is distributed under the terms of the Creative Commons Attribution-NonCommercial 4.0 License (<https://creativecommons.org/licenses/by-nc/4.0/>) which permits non-commercial use, reproduction and distribution of the work without further permission provided the original work is attributed as specified on the SAGE and Open Access pages (<https://us.sagepub.com/en-us/nam/open-access-at-sage>).

Various studies aiming to support the “aging well” of older adults are being carried out using ICT and robotic technology under these circumstances. A typical example of the efforts is an approach called “ambient assisted living (AAL),” which supports independent living of older adults as much as possible by applying technology such as internet of things (IoT).⁴⁻⁷ Systems that provide services in cooperation with various sensors and robots connected through networks, robotized smart houses, and remote monitoring services using sensors are typical examples.⁸⁻¹²

In these systems, the required information is obtained using sensors, and the action or output is delivered through robots or presentation devices. Thus, the quality of the output depends on the types and setups of the sensors used. For example, it is necessary to determine the state of the living environment, such as room temperature and humidity, and the behavior of the older adults themselves, to learn their living conditions.

In most cases, however, the detectable information is only related to the external state of the older adults or the state and behavior exhibited on the outer surface and provides only speculation about their internal state. Such input is therefore imperfect and insufficient, considering the purpose of learning about their physical and mental health. The common method humans use to learn the inner state of another person is a voice call and perceiving or hearing responses to the call. Here, our idea is to use this strategy to design a robot that speaks to older adults and senses their responses to obtain more accurate information directly related to their physical and mental health.

This can be said to be an active sensing method in the sense that the person’s inner state is known through output reactions for the voice-call input. It is understood that the content of the voice-call is an important issue when obtaining the necessary information to estimate the physical and mental conditions of older adults. It is desirable and essential to make the interaction a form of dialogue that continues as long as possible to acquire as much information related to the person’s conditions as possible. Furthermore, it is also important to consider the kind of assistance that could be provided based on the information obtained, including voice-calls, to help maintain the physical and mental health of the older adults.

This research can be regarded as an AAL system in which a voice-call is performed as one of the major ways of sensing conditions, as explained above. This paper describes the results obtained at an earlier stage of the research and the method adopted to extend the dialogue for longer periods.

The structure of this paper is as follows. Basic concept of the system and development of the system are described after introduction section. Then, pairs of methods and results for three user studies are explained in a chronological order, since we had refined the experimental system step by step, in such a way that the issues revealed at the time of

each experiment were being improved. In the discussion section, user studies 1, 2 and 3 are arranged and discussed in parallel manner, so as the comparison between the methodologies and results can be understood easier. We will summarize the findings and implications for future studies in conclusion section.

Basic concept of the system

Background of the proposed system

Naturally, the greatest concern in assisting older adults is about their health problems. The research and development activities related to AAL systems deal with this issue by utilizing various IoT sensors and robotic devices. Some examples¹³⁻¹⁵ include monitoring systems¹⁶ that can activate calls to care centers in case of emergency^{17,18} or that can notify them about wandering dementia patients.

The health condition of a person in daily life can generally be estimated from the irregularity in the continuation of a regular life pattern; thus, it is meaningful to pay attention to the temporal transition of sensor information reflecting the life pattern of older adults living alone. The time to get up or the time to go to the bathroom are some examples that indicate life patterns. Another such example is the number of times the person goes to the bathroom, which correlates with the amount of water intake¹⁹ and is considered to be a guide for the intake of necessary water.

Sensing such externally appearing behavior can be used as an index for the physical and mental health of a person, although it is only indirect information. From this point of view, we propose a novel idea, which is to positively and directly sense the physical and mental health of the person by voice-calls, using a robot, and observing their response to the voice-calls, in addition to the conventionally available sensors’ information. It should be noted that the reason for using a real robot, instead of the voice-calls from a simple speaker, is to promote the feeling of affinity and presence of a partner by the older adults, especially those living alone, by talking to an entity that is substantial and to inspire long-term use through such feelings.

The most important issue when using voice-calls in sensing the state of a person is to prepare appropriate content suited to obtain the necessary information regarding the health status of the person. It is desirable to continue the interaction of voice-calls and responses, that is, the dialogue, as long as possible to acquire as much information as possible. Here, we assumed that the contents or a basic scenario of the voice-call could be determined in a fixed manner, since the purpose of the voice-call in this case is to obtain information regarding the physical and mental health of the older adults. However, certain considerations were given, such as changing some parts of the voice-call contents according to the season, time of the day, weather, etc.,

to make it possible to continually use the content for a long period of time with less discomfort.

In this research, we have defined the term “scenario-based dialogue” as a set of dialogue consisting of predetermined statements in such a way that the next statement from the robot is not unnatural, regardless of the utterance made by the user in response to the first statements made by the robot. This method has the features that the dialogue can be established even when the pronunciation of the older adult is not clearly recognizable, or that the contents of the robot’s speech can be carefully prepared in advance. Regarding such scenario-based dialogue, we have already succeeded in making the dialogue to continue up to two user’s turns.¹⁸

Along with the content of the linguistic information spoken by the older adults, the reactions and non-verbal information during the interaction of the older adults with the robot are also considered effective in estimating their health condition. Research on estimating diseases from voice has been conducted. Stress level,²⁰ arousal level,^{21,22} and in the case of depressed people, the speed of the voice,^{23–25} volume,²⁶ acoustic information,²⁷ and the number of pauses^{25,28} have been studied as characteristics that are thought to be key to estimating mental state.

Voice volume and tone of the voice are the examples of non-verbal information. The content of the response to the robot’s voice-call regarding specific topics, such as meal, sleep, confirmation of medication, as well as the length of the response, the reaction speed, etc., are also considered to be useful information for estimating the life pattern and physical condition of the older adults.²⁹ Caregivers usually talk to older adults to check their physical condition. If caregivers find some abnormality in the older adults, they give specific countermeasures and advice. When engaged in dialogue, if someone asks an inconvenient question, the other person may reply after a short delay, or the volume of the reply may be lower than usual. Moreover, emotional information extracted from the voice of a user includes these acoustic characteristics.^{30–32} Human emotion is expressed not only toward humans but also to robots; this emotional information is used for the design of robots’ behavior.^{33–37} Therefore, we hypothesized that the emotional information in the user’s response to the robot’s voice-call is key to estimating the user’s life patterns and physical conditions.

Thus, by using dialogue, it is possible to proactively interact with older adults utilizing the robot’s voice-calls, actively sense their daily behavior and status through voice responses, and accurately estimate their health condition. Furthermore, the robot can make recommendations to the older adults that are useful for maintaining and improving their health, by responding to users based on the estimated health condition of the older adults. Additionally, we considered it useful for older adults to think about the

response to the robot’s voice-call themselves to maintain their cognitive function.

As mentioned above, it is possible to obtain more information regarding the health condition of older adults if the conversation with the robot is maintained for a long period. It is also desirable to build an intimate relationship between the robot and the human as much as possible when the robot makes recommendations for maintaining the older adult’s health using a voice-call.

Positioning of this study

The development of an AAL system with a dialogue robot that acts as a dialogue partner for older adults consists of various technologies such as sensor hardware, electronics, data processing, robotics controls, etc. Technologies related to dialogue management and control must solve elemental problems by developing a: (I) method to keep conversation between an older adult and a robot for longer period of time, (ii) a method to prolong the older adult’s interests and to maintain engagement with the robot, (iii) the contents of the robot’s voice-call suited to obtain the necessary information regarding the person’s health status, (iv) a method to effectively extract the person’s health status based on the utterance of the person³⁸ along with the output data from various sensors, (v) the contents of the robot’s voice-call to give appropriate recommendations for promoting health-friendly activities for the older adult. It should also be able to identify character differences among persons and find ways to deal with these differences so as to individualize the system applicable to the practical AAL application. In this paper, the methods outlined above, specifically the method to keep conversation between the older adult and the robot going for a longer period of time, has been studied as the first step for developing the system explained earlier. A method for estimating a person’s rhythm of speech and end-of-turn utterance has been proposed, and its effect has been experimentally evaluated as the preliminary study of this first development step.

When considering the case of dialogue between people, it is known that the timing or rhythm of turn-taking is important for continuing the conversation comfortably, while maintaining a good mutual relationship.^{39,40} For this reason, we have decided to introduce a method to make the dialogue continue as long as possible, by controlling the robot’s turn-taking and the timing of the start of a voice-call, based on the utterance pattern of older adults. We have also studied whether the method is effective in the dialogue between humans and robots, especially in the cases where multiple turn-takings (more than three turns) are included.

In this study, we investigated the user’s reaction to the voice-call of the robot and the possibility of controlling the turn-taking process by using the blank time after the users response to the voice-call as a variable in the study. The

main purpose of this study was to collect fundamental data regarding the user's reaction to the robot's voice-call, and therefore the contents of the user's utterance were not the focus of our the analysis. Therefore, we did not apply voice recognition techniques to the user's utterance, and a simple scenario-based method with mostly fixed content has been employed. A machine learning technology utilizing neural networks was applied as a strategy for determining the appropriate turn-taking timing.

In this paper, we describe the preliminary experiment regarding the system based on the above-mentioned concept, the results on the user's reaction, and the analysis of data, such as the duration and number of blanks between the dialogues obtained through the experiment. We also provide an outline of the experimental system, results of the preliminary experiments, the user's reactions, including the analysis of data obtained through experiments, such as the duration and number of blanks between the dialogues.

Development of a scenario-based dialogue system

Scenario-based dialogue system

Figure 1 shows an example of scenario-based dialogue with a robot and a user. The rectangle bar represents the sound of speech for the robot and the user. The content to be spoken by a robot is decided in advance, since this system is a scenario-based dialogue. As a first step, we constructed a robot system to control the timing to begin speaking. For privacy reasons, it is better not to use facial data for turn-taking. Moreover, in the case that older adults cannot speak clearly, the accuracy of the speech recognition results may not be high enough to allow the robot system to take turns. We started by developing a simple system as a first step.

To develop a scenario-based dialogue system, we used information from the on/off microphone for turn-taking. When the system detected a soundless segment, the section could be (1) the user's breathing during his/her reply or (2) the end of his/her reply.

Steps for developing the robot system

We developed the trial production, user study, and improvement in steps. In the development, we used a scenario in which the user and robot dialogue at home. We report the development process of a scenario-based dialogue system with an automated turn-taking function, which involved three steps of development and user studies in this paper. The paper is constructed as follows: a scenario-based dialogue user study to collect data for prototyping (User Study 1), a prototype of a scenario-based dialogue system with a silence-based turn-taking system (Development and User Study 2), and an update of the automated scenario-based dialogue system and an attempt to individualize turn-taking (Development and User Study 3).

First, we conducted a user study to investigate the reactions of users to the robot in scenario-based dialogue. In the user study, the timing at which the robot voice-calls is remotely controlled by an operator. Based on the results of the user study, we then constructed a prototype of a scenario-based dialogue system that can automatically control the timing of the start of a voice-call, followed by the user study, again using the prototype system. All of the dialogue user studies were conducted via video chat communication system Zoom as a video chat software (Zoom is a trademark or a registered trademark of Zoom video communications, Inc.) (see Figure 2)

The robot we used this study was BONO-06 (Figure 3,^{41,42}) and the robot was displayed on the screen of the PC/tablet of a user. The voice of the robot is synthesized in advance and played back at the timing of the speaker change. We used ReadSpeaker⁴³ for speech synthesis. As for the voice parameters, a high, childlike voice was employed to give users a positive impression. Before a dialogue segment begins, the volume of the sound was adjusted for each user. After User Studies 1, 2, and 3 were finished, we brushed up on the parameters of the robot's voice-call based on the opinions of our study's participants. Specifically, the speed of the robot's voice-call was updated. We updated the parameter of the sound of robot's voice-call depending on the step of the development of this system. To

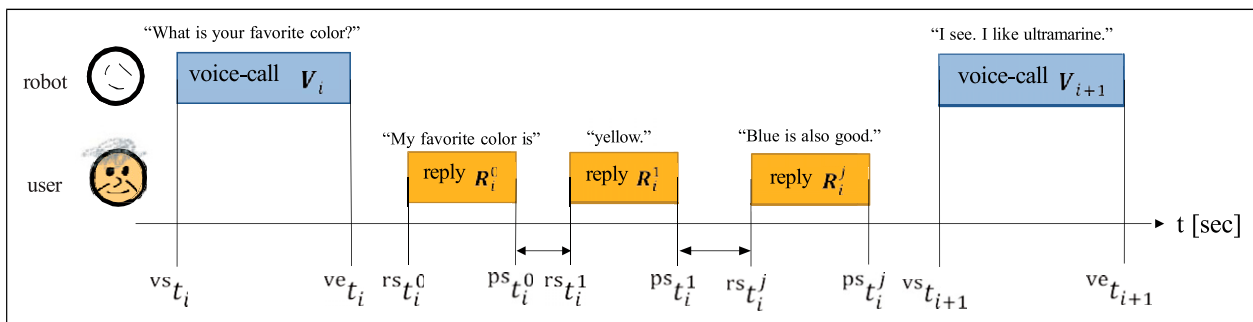


Figure 1. Overview of scenario-based talking between a robot and a user.

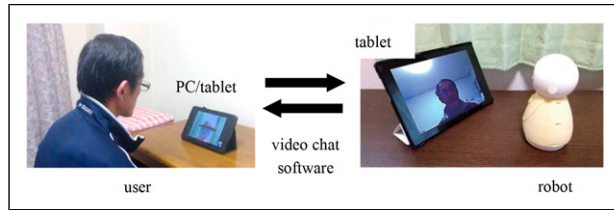


Figure 2. Scene of the user study of scenario-based dialogue and turn-taking with video chat software.

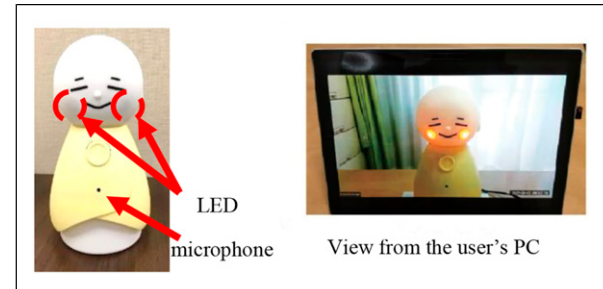


Figure 3. The appearance of the robot BONO-06.

help older users hear the robot's voice-call more clearly, we updated the parameters of the sound synthesis of the robot's voice-call depending on the step of this system's development based on user feedback. We updated the parameters of the sound for the robot's voice-call with step-by-step approach. To realize the actual situation that the robot is set as if in a house that appears on the screen of video chat software. Also, the sound of the user is output from the PC which runs the video chat software and was recorded by a microphone attached to the robot system.

The studies described in this paper were conducted in accordance with the research ethics committee of RIKEN (No. Wako3 2019-31). Written informed consent was obtained from the participants in accordance with the Declaration of Helsinki. All participants provided informed consent and the consent was written.

Scenario-based dialogue user study to collect data for prototyping

User Study 1: Survey of users' reactions to the robot

Aim: To collect data on pause duration and the number of pauses during the robot's voice-call and to build a prototype of a scenario-based dialogue system, a user study was conducted in which the user and the robot had a scenario-based dialogue.

We also conducted a questionnaire to qualitatively evaluate the naturalness of the content of the scenario-based dialogue.

Method: The timing of the starting robot's voice-call was controlled by an operator. The procedure is as follows (Figure 4).

1. The robot starts the first voice-call V_i , ($i = 0$).
2. The robot waits for the user's reply R_i^j to the voice-call V_i .
3. After the user's reply R_i^j is over, the robot starts the next voice-call V_{i+1} .
4. If the robot's voice-call V_{i+1} and the user's one more reply R_i^{j+1} overlap, the robot stop the voice-call V_{i+1} and waits for the user finish his/her reply R_i^{j+1} .

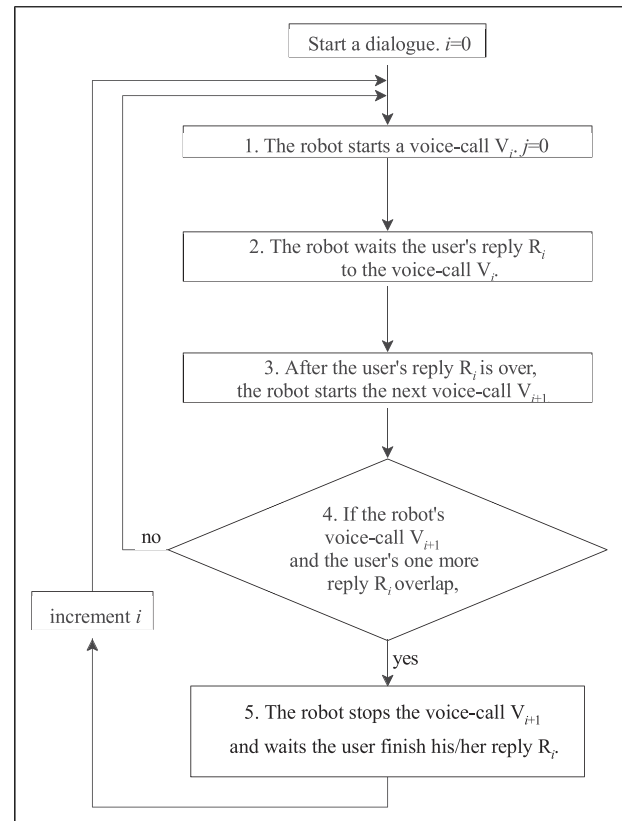


Figure 4. The flowchart of the system for user study 1.

The scenario used in User Study 1 is shown in Table 1. Table 1 shows the whole dialogue that was performed in User Study 1. For each participant, in order from the top, the robot gave these voice-calls one by one. After a speech was given, the robot waited for the user's reply. Then, the robot spoke again. The robot spoke 17 turns in total. The dialogue for one user for one time took about 5–10 min. The Type column in Table 1 represents the classification of the content of the robot's voice-call. This classification was used to analyze the user's responses, which are thought to differ depending on the content of the robot's voice-call. Table 2 describes the type of classification of the voice-calls.

Table 1. Scenario script for the robot and the type of the contents of the voice-calls (User Studies 1 and 2).

Turn no.	Type	Voice-calls
1	S	It's nice weather today.
2	S-R	I would like to take a walk.
3	OQ	Yes. If you go for a walk, where would you like to go?
4	OQ	Now that I have more time at home, I'm thinking of starting a new hobby. What is your hobby?
5	OQ-R	I also would like to try that.
6	S	It's still cold in the morning and at night, so please wear warm clothes.
7	S-R	The brighter the color of the clothes you are wearing, the happier you feel.
8	OQ	What is your favorite color?
9	OQ-R	Is that so. I like ultramarine blue.
10	S	By the way, the number of warm days has increased.
11	S-R	At times like this, I would like to eat something warm.
12	OQ	What is your favorite hot food?
13	OQ-R	I see, I would like to eat it too.
14	S	2021 is already February/May.
15	S-R	I would like to be a robot that can play an active role this year.
16	OQ	By the way, what are your goals for this year?
17	OQ-R	Yes, let's have a good year with each other.

Table 2. Type of the voice-call of a robot.

Type	Contents of the speech
S	Information presented by a robot
S-R	Information related to the previous information presented by a robot
OQ	Open-Question
OQ-R	Information about the previous open-question

We also conducted another dialogue using a chat system to compare these evaluation items. The chat system has the three following functions: speech recognition, getting a reply, and speech synthesis. The chat system recorded the user's speech for 3 seconds and the recorded sound data were converted into text data. The system then received a reply generated by a chat dialogue system, Katarai API*, based on the speech data, and synthesized sound data were output as the robot's voice-call. We assumed a short reply of a user to the chat robot, and the duration of the speech recognition was 3 seconds not to delay the timing of the robot's voice-call. The chat dialogue system used in User Study 1 outputs a reply regardless of the length of the input. One keyword is thought to be enough to output the robot's reply. The robot asked some open questions prepared in advance in case the speech recognition is not successful. After the number of the robot's voice-calls reached the number of the scenario-based dialogue lines, the dialogue finished with the robot saying, "It's about time to end. Let's close this dialogue now."

To evaluate the naturalness of the content of scenario-based dialogue, we conducted two types of methods to determine the content of each dialogue. At first, the user had

a dialogue with the scenario-based dialogue system. Then, the user had another dialogue with the chat-based dialogue system. Half of the participants took a reverse order for counterbalance. The user answered the questionnaire after each dialogue experiment of the scenario-based dialogue and the chat-based dialogue. The naturalness of the scenario-based dialogue, the enjoyableness of the dialogue and the level of desire to continue the dialogue were evaluated by 5-point Likert scale⁴⁴ questionnaires after each dialogue.

Four healthy older adults (from 60 to 80 years old) of two men and two women participated. The participants were accustomed to digital devices such as smartphones and laptop. The users were instructed to answer to the robot's voice-call in as much detail as possible before starting the user study.

We focused on silent segments for voice-call interactions from robots to humans. There were two cases of silent segments. First, there was a silent segment in the middle of the user's talking. Second, a silence segment after a user said one sentence. In this study, we distinguished these silent segments caused by breathing in the middle of the user's reply from another soundless segment after the end of the

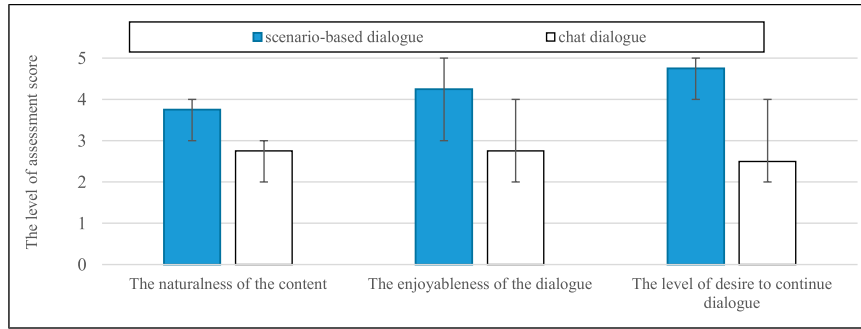


Figure 5. The level of assessment score: scenario-based dialogue versus chat dialogue in User Study 1.

user’s reply. We define the silence segment caused by breathing in the middle of talking one sentence as the “pausing segment.” The silence segment after one sentence was not the pausing segment.

During this user study, the robot system was connected to the microphone and recorded the time when the sound of the user’s utterance started and the time when the silence segment starts. After the user study, the recorded time of the silence segments were manually divided into the pausing segments and the end of speech segments based on what the user said in the dialogue. This manual work was done by one researcher according to the protocol of the work below. Using transcribed text data of reply of the user and the sound data of the user’s reply to the robot recorded during User Study 1, we marked the timing of when the sound arose and when it disappeared, calculating the duration of the pauses accordingly.

Result: First, we show the results of the evaluation of the naturalness of the scenario used in this user study. Figure 5 shows the results of the questionnaire in the first experiment. The users seemed to feel that the scenario-based dialogue was more natural than the chat dialogue system, in which voice recognition often failed. However, the chat-based dialogue system in User Study 1 was preliminary, therefore a more natural chat-based dialogue system needs to be compared to evaluate the naturalness of the contents of the scenario-based dialogue.

Figure 6 shows the duration of pause and non-pause pausing segments for each participant. The duration of blank during pausing was about 0.5 s, and the duration of non-pause segment was longer than twice the length of the duration of pause. Comparing the variances of the distribution of the blank data, the variance of the blank in the non-pause segment was larger than that of the blank of pause segment.

Next, we counted the number of blanks that are included in the users’ reply to one robot’s voice-call (between the robot’s voice-call V_i and the next voice-call V_{i+1}). For each type of blank (blank of pausing and other

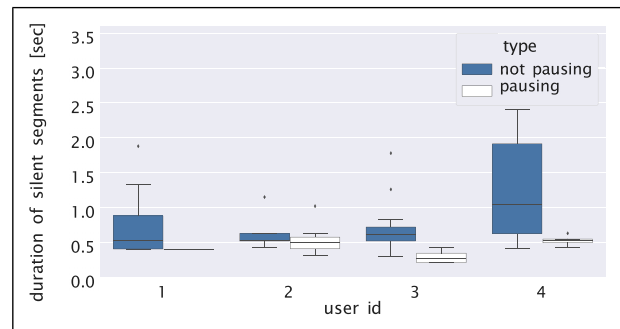


Figure 6. Duration of the pausing segments and the end of the speech segments.

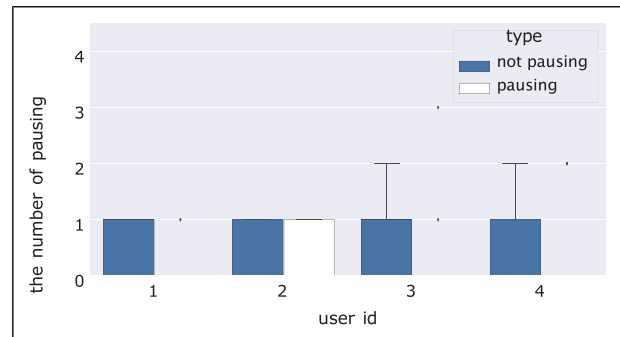


Figure 7. Average number of blanks during a reply to the robot’s voice-call according to the types of blank in the 1st user study.

blanks), Figure 7 shows the average value of the number of blanks included in one user reply. Overall, for each user, the blanks were about 0–1 time. When the pausing segments and the other blanks were combined, one reply contained about two to four blanks. We considered that it was good to wait for two to four times of blanks during a user’s reply. Since the number of blanks except for pausing was the same as the number of sentences included in the user’s reply to one voice-call of the robot, if this number was large, it suggests that the user has spoken to the robot many times.

Consequently, the number of blanks may change depending on the user's attitude toward the dialogue with the robot.

Sometimes the robot's voice-call and the user's reply overlapped because the user started speaking, trying to break a silence. There were five cases in which the user attempted to break a silence, and the duration of the silence was about 1.3–2.5 s. To deal with such a situation, the robot system would need to have a function that detects overlapping and stops the robot's voice-call if an overlap is detected.

As a conclusion of this user study, the outline of the results of this user study and key marks are listed as follows. The following results become indicators to encourage a human to speak at a good tempo.

In this study,

1. the blank for pausing was about 0.5 s,
2. the duration of the blank before the user's speaking was from 1.3 to 2.5 s.

Prototype of a scenario-based dialogue system with a silence-based turn-taking system

Silence-based turn-taking system

We used the same experimental setup of the User Study 1, comprising a robot and remote meeting system. The difference was that the turn-taking was automated based on the detection of silent segments by processing the sound level.

The previously proposed methods for voice interaction systems of turn-taking generally employ voice recognition results^{45,46} the user's gesture or gaze,⁴⁷ or prosody information^{48–51} to detect the end of turn.

For turn-taking, the four states of the user's speech—"silence state, utterance start, utterance in progress, and utterance end"—were detected based on the time-series data of the sound level of the voice of the participant, which is acquired by the microphone. We defined four modes of speaking of users: "pausing," "start talking," "talking," and "start pausing."

- "Pausing" is the state in which the sound level is lower than the threshold value. In this state, the user is assumed to listen to the robot's voice-call or pause during his/her speech.
- "Start" is the moment when the sound level increases. In this state transition, the user is assumed to start speaking.
- "Talking" is the state in which the sound level is higher than the threshold value. In this state, the user is assumed to speak to the robot.

- "Pause" is the moment when the sound level decreases. In this state transition, the user is assumed to stop his/her speech.

In addition to the automation of turn-taking, we improved the following based on the users' comments. To help turn-taking, the color of the LEDs on the robot's cheek was changed according to the user's speaking state. However, to determine how the user reacted, the user was not informed in advance of the robot's color change.

The speed of the robot's voice-call in User Study 2 was set slower than in User Study 1 because some users felt it difficult to hear the robot clearly in User Study 1. Moreover, the sound of the robot was output directly from the video chat software's speaker to allow users to hear the robot's voice-call clearly.

User Study 2: Prototype experiments of an automated scenario-based dialogue system

Aim: We conducted a second user study of dialogue using the constructed system and investigated the reactions of users and some problems in the dialogue. We increased the number of participants with attributes similar to those in User Study 1 and conducted a dialogue experiment. We investigated problems and challenges that may arise during the dialogue and analyzed the data obtained in the dialogue. Changes in participants' reactions (pausing duration, the number of blanks, the amount of talking, etc.) according to the content and the naturalness of the content of the robot's voice-call were also investigated.

Method: Eight healthy older adults (from 60 to 80 years old) of three men and five women participated in this user study. Four adults also participated in User Study 1. The four new participants were healthy older adults, same as the previous four participants, and they were also accustomed to digital devices such as smartphones and laptops. The specific procedure for the robot voice-call system is as follows (Figure 8).

1. The robot starts the first voice-call V_i , ($i = 0$).
2. The robot waits after detecting the start of a user's reply R_i to the previous voice-call V_i .
3. The robot waits for 1.5 s after the system detects the user's starting pausing of the user.
4. After the silence segment continues for 1.5 s, the robot starts the next voice-call V_{i+1} (back to 2nd procedure).
5. If the robot detects the user starts speaking again within 1.5 s after the robot starts waiting, the system waits for detecting the starting pausing of the user (back to 3rd procedure).

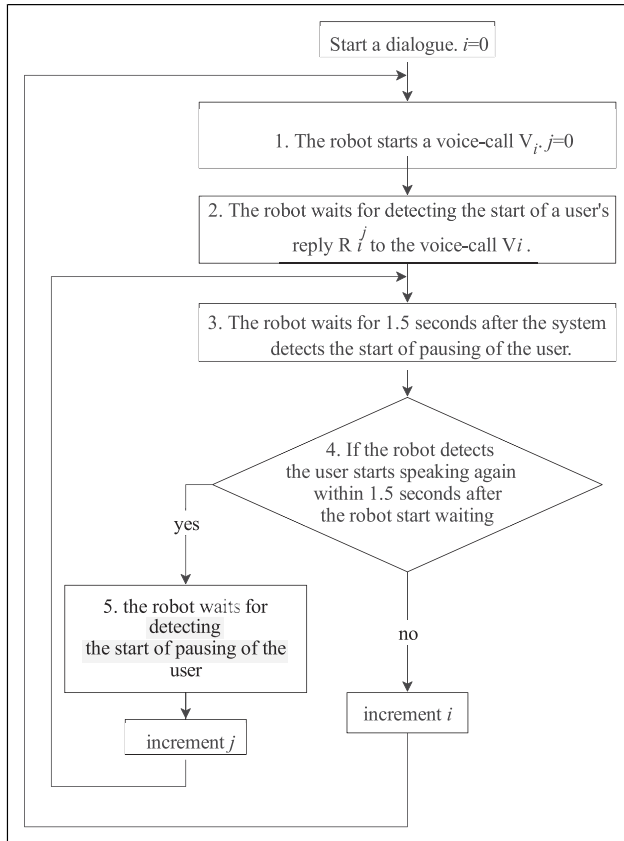


Figure 8. The flowchart of User Study 2.

The users were instructed to answer to the robot's voice-call in as much detail as possible before starting the user study. The same questionnaire as in User Study 1 was conducted.

We decided that the condition of waiting time for the robot to start a voice-call was that the silence segment lasts for 1.5 s; the number was based on the result of User Study 1. Table 1 shows the dialogue scenario in this study. The type represents the kind of utterance applied to the voice-calls which are defined in Table 2. OQ represents the abbreviation of the open question. S represents the abbreviation of the statement. OQ-R represents the response to the user's answer to the previous open question by the robot. S-R represents the response to the user's comments to the previous statement by the robot. The voice-call row represents the voice-call script from robots to older adults. For example, turn number 1 indicates that the robots starts to talk with a generic greeting with type "S" and says to older adults, "It's nice weather today."

Result: To show whether the turn-taking succeeded in User Study 2, Figure 9 shows the number of times the robot's voice-call and the user's speech overlapped, and the number of times the robot started voice-call before the user's speech was completed. Using the constructed

turn-taking system, the robot system usually succeeded in turn-taking with users.

The results of the questionnaire were good (Figure 10). The evaluation results for the scenario-based dialogue with the automated turn-taking system were generally higher than the standard score of 3, suggesting that users enjoyed the dialogue and were motivated to continue the dialogue in User Study 2. However, there were some problems when the robot's voice-call and the user's speech overlapped. When the robot stopped its voice-call after the robot and a user overlapped, the user also stopped his/her speech and the silence continued for a while. The maximum length of the silence was about 40 s. The other problem was that the robot was not able to distinguish between the user's interrupts and back-channeling. When the robot scenario was long and the user's response to it was judged as an interruption, even though the response was back-channeling, the robot stopped its voice-call and a silence continued. The user looked to feel the robot stopped because of some accidents.

We summarize the results of the users' reactions, including the blank and speech length data except for the data when the problem described in the previous paragraph occurred (Figures 11, 12, and 13). The reaction of the users was thought to differ according to the type of robot's voice-call. When the robot's voice-call was an open-question that prompted the user to think freely, the blanks during the user's reply was thought to be more than the case when the robot's voice-call was not open-question. We divided the user's reactions according to the type of the robot's voice-call: open-question or non-open-question.

Figure 11 shows the duration of the blank of pausing for each user. The blank of pauses during a reply to the robot's voice-call which is not open question was about 0.5 s. The duration was close to the result of User Study 1. The blank of pauses during a reply to the robot's voice-call which is open question was almost the same. However, for some users, the variance value of the blank duration of open questions was longer than the blank of non-open questions. Figure 12 shows the number of pauses for each user. The number of pausing segments during a reply to one voice-call involving "non-open question" was up to six times, about one to five times. By contrast, in the case of open questions, the number of pause blanks differed according to the users.

Based on Figures 11, 12, and 13, the users were thought to require more time to think and speak if asked open questions compared with non-open ones. The word count and number of pauses of the users' answers to the robot's open questions (OQ) are thought to be higher than those of the answers to the non-open questions (NOQ) asked by the robot. Also, the pausing duration is thought to be longer

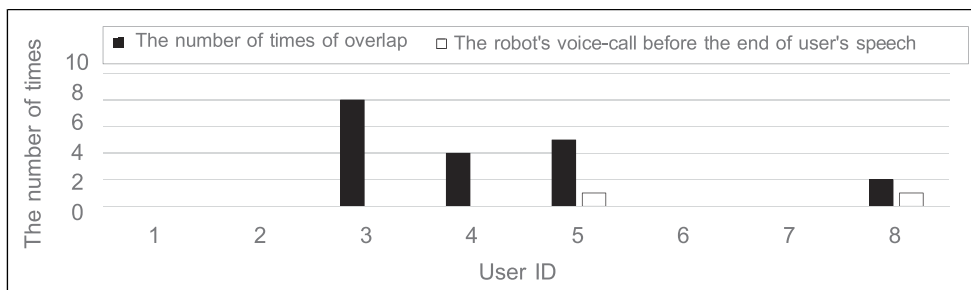


Figure 9. The number of overlaps, and the number of times that the robot started a voice-call before the end of the user's speech according to each user in User Study 2.

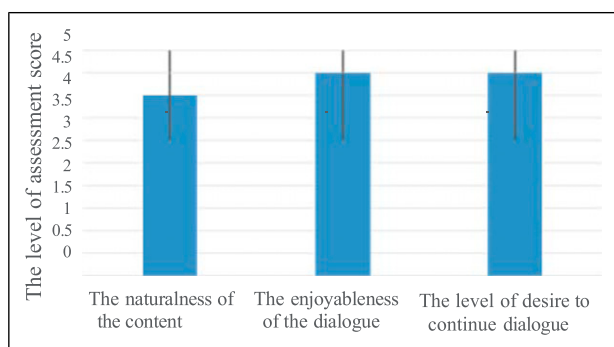


Figure 10. The level of assessment score in User Study 2.

because the user has to think about what to say in their reply instead of simply answering yes or no.

The speech duration, the word count of the speech, the pause duration, and the number of pauses were summarized according to the type of the robot's voice-call (Figures 14 and 15). As for speech duration, word count, pause duration, and the number of pauses, Welch's test⁵² was used because these two distributions were independent samples and these were unequal population variances. In each case, the p -value is smaller than 0.01, and there is a significant difference between the average values of the two groups. The amount of the users' speech, the number of characters of the users' speech, the pause duration, and the number of pauses during the responses to the open questions tended to be longer and higher than these features of users' responses to the non-open questions.

The number of pauses and the pause duration of the responses to the non-open questions had less variation than those to the open questions. From examining these findings, it is revealed that the system does not necessarily have to consider individual differences regarding pause duration during the user's speech in the case in which the robot's voice-call is a non-open question.

On the other hand, the number of pauses and the pause duration had greater variance in the case of the open questions. Some users are thought to spend more time thinking about what to answer depending on the content of

the question. For example, depending on whether the user thinks about the future or is reminded of the past, the speech duration and the pause duration change, and this may differ depending on the individual's personality or way of thinking.

Regarding the users' responses to the open question, it seems to be necessary to further examine the influence of individual differences and the content of the robot's voice-call on the pause duration and other features of the user's responses.

Update of the automated scenario-based dialogue system in an attempt to individualize the turn-taking

Based on the results of User Study 2, we updated the scenario-based dialogue system. We also built a method for individualizing the timing of starting the robot's voice-call. First, we explain the method of individualizing the turn-taking system. Second, we explain User Study 3, which was conducted to confirm the operation of the dialogue system and check the reactions of users and the effects of the dialogue system.

System of individualizing the timing of turn-taking of robot

How to individualize the timing of turn-taking: This section describes a method of individualizing the timing of the turn-taking of a robot. In a dialogue with a user and a robot, we think it is better that the robot individualizes the timing of the start of a voice-call according to the character of the user to regularly continue the dialogue. For example, the robot should wait longer and start the voice-call after the user ends his/her speech if the user usually talks slowly. If a user usually talks with a good tempo, the robot should start the robot's voice-call at a faster time. We built a system to individualize the timing of turn-taking by controlling the timing to start the robot's voice-call, based on the prediction of different durations of blank for each individual. Two subsystems

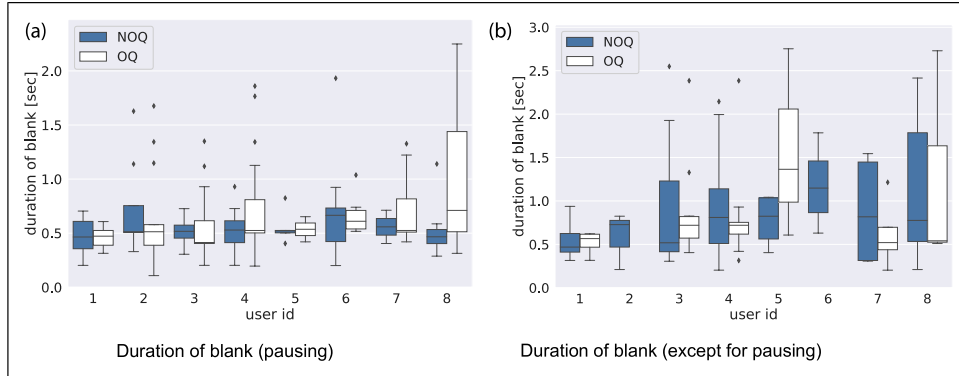


Figure 11. Duration of blank according to the type of robot's voice-call.

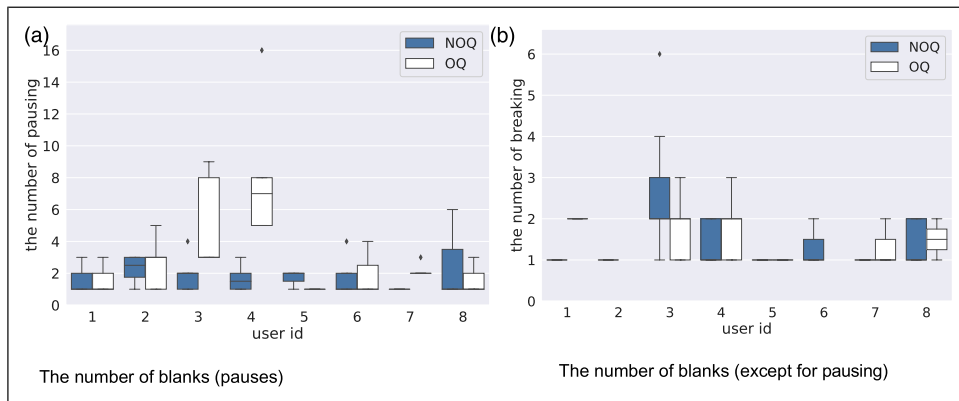


Figure 12. The number of blanks according to the type of robot's voice-call.

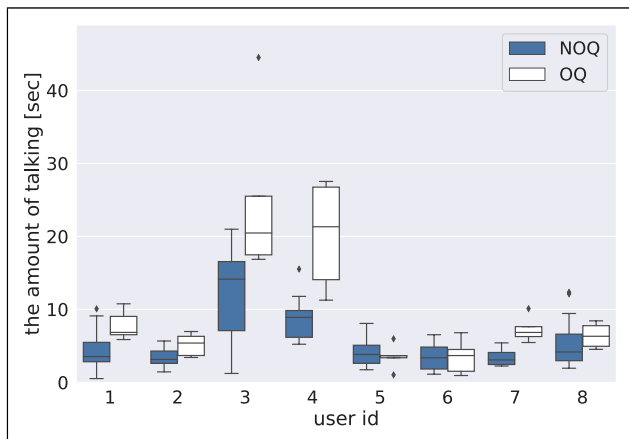


Figure 13. The amount of speech of users according to the type of robot's voice-call.

System of predicting the duration of the blank and the number of the blanks: In this section, we describe how we build a predictor framework that predicts the timing to respond to a user based on the number and duration of blanks in a user's response. We also describe how we developed such a machine learning model. Specifically, the system estimates user reactions (blank duration and the number of pauses) that differ from person to person. The system estimates the maximum number of pauses and the duration of pausing during the user's reply to each voice-call of the robot. However, in real life, sometimes it is difficult to acquire user-specific speech data in advance for parameter tuning of an individualized dialogue system. Therefore, we employed a method for estimating the reaction of a user using online learning with the state in which the parameters are not tuned for each user as the initial state.

comprised the individualized dialogue system: a system of predicting the duration of a blank and the number of blanks for each user and a system of controlling the timing of the start of the robot's voice-call based on the prediction.

A method for controlling the timing of the robot's voice-call based on the predicted result will be described. First, in this method, the type of the contents of the voice-call was used to predict the user's blank space. This is because in User Study 2, the user's blank duration tended to differ

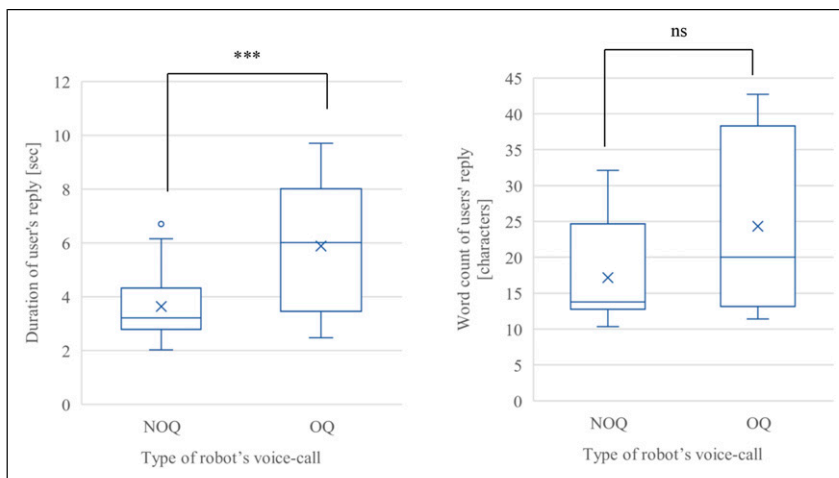


Figure 14. The difference in the duration of users' replies (OQ: open question vs. NOQ: non-open question) *: p -value < 0.05, **: p -value < 0.01, ***: p -value < 0.001, ns: p -value > 0.5.

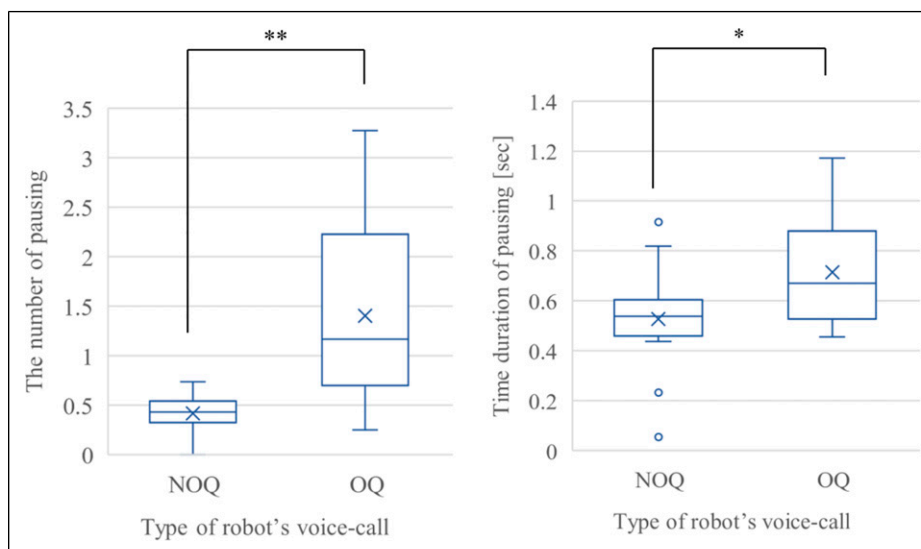


Figure 15. The difference in the number and duration of pauses (OQ: open question vs. NOQ: non-open question) *: p -value < 0.05, **: p -value < 0.01, ***: p -value < 0.001, ns: p -value > 0.5.

depending on the type of robot's voice-call (whether the voice-call was OQ (open question) or not). Second, the reaction of a user is thought to be affected by the number of voice-calls about a topic because the more times the robot voice-calls about a topic, the longer the user has to think his/her reply. Further, except for some examples, the variance of the blank duration or the number of blanks for each user was small in User Study 2. For blank prediction, another clue is the average value of blank duration and the number of blanks for each user. We supposed that the content (topic) of the robot's voice-call affected the user's reaction; however, it takes a longer time to learn various interests of users in detail by online learning. Collecting more individualized

data during a dialogue is an issue to be addressed for the next step.

Based on the data described above, online learning was performed to predict the user's blank time and number of times. In this method, we employed a neural network (NN) as a framework for learning the reactions of individual users. The aim of using NN's was to address situations that had not been experienced in the past by the generalization capabilities of NN. We regarded the prediction of the users' reactions as a problem of predicting the duration and number of pauses, not as a classification problem. Therefore, we employed NN instead of a support-vector machine (SVM⁵³) or decision trees.

Generally, overfitting should be avoided, however, we do not address overfitting at this point because the values that are normally considered outliers could be individual characteristics.

The data used for learning are explained next. When the robot makes the i -th scenario voice-call V_i , it predicts the maximum number of blanks and the blank time included in the user's response. Here, it is assumed that the user's response contains multiple blanks. Of the user's responses to the robot voice-call V_i , the user's response after j blanks are detected is R_j^i (see Figure 1). In what follows, we describe how our NN model consists of input/output parameters that are key to predicting both pauses and blanks. The input vectors for the NN were as follows:

- The type of voice-call: binary value in which the voice-call is OQ or not.
- Number of topics: How many times a topic is called.
- Number of pauses: The number of blanks confirmed so far for one reply.
- Average value for each user: blank time during response to robot's OQ voice-call, blank time during response to robot's non-OQ voice-call, and the number of blanks in reply to OQ and NOQ voice-call.

The output vectors—that is, the data to be predicted—were:

- Next maximum blank time,
- How many more blanks will occur.

Using these data, we constructed a predictor that predicts individual reactions.

From the state that the predictor was trained using the data of multiple users, the parameters of the predictor were tuned using the individual participant's data. This is to get closer to the parameters that match the individual data faster.

At the beginning of the dialogue user study, the user practices dialogue with the robot to obtain the average value of the blank duration and the number of blanks of the user. In this practice, the timing of starting the robot's voice-call is fixed, and online learning is not performed. The user and the robot have a dialogue including a few voice-calls (the scenario that was used in User Study 3 is shown in Table 3) and the average value of the blank duration and the number of blanks of the user are calculated based on the user's speech data obtained. After practicing, a new dialogue starts, and online learning starts at this point. That is, the user's reaction is predicted using the blank predictor, and the timing of starting the robot's voice-call is controlled based on the prediction. Immediately before starting the next voice-call, the system predicts the number of blanks and blank duration included in the next user's response based on the type of the next voice-call, the number of topics of the

voice-call, and the average value of the blank duration and the number of blanks that were obtained in the practice.

How to individualize the timing of start a voice-call: The timing of the robot's voice-call is controlled as follows (also see Figure 16): After the next voice-call is determined,

1. If the predicted number of blanks is zero, the system starts the next voice-call after the end of the first blank is detected.
2. If the predicted number of blanks is not zero, the system waits for the predicted blank duration after detecting the start of the blanks.
3. When the predicted blank duration has elapsed, the system starts the next voice-call.
4. If the user starts his/her speech again before the predicted blank duration elapses, the next blank duration and blank time of the user are predicted again, and the process returns to 1.

In the actual experiment, the system learned the relationship between the input/output data based on the reaction data of all the participants in the User Study 2, and this was set as the initial state.

User Study 3: Individualizing the timing of robot's voice-call and surveying user's reactions

Aim: The individualizing system was implemented in our dialogue system, and we conducted a dialogue experiment. We confirmed the operation of the dialogue system and checked how individualizing affected the user's reaction and the amount of talking.

This dialogue system was updated with some points based on feedback from User Study 2.

Method: We increased the number of participants in User Study 3, and they had a dialogue. We had two ways of controlling the robot. One was the automated dialogue system, which was updated based on the results and feedback of User Study 2. The other was the method with an individualized turn-taking system described in the previous section. The scenarios used in the dialogue were almost the same (Table 4). The scenario was designed to directly or indirectly ask about the user's health status and to implicitly recommend behaviors that would improve the user's health. Topic 1 directly asks health condition. Topic 2 asks interest in exercise, which implicitly recommend walking. Topic 3 asks memories of the recent past, namely, yesterday, which reflects memory functioning. Topic 4 asks meal to check nutrition status. Topic 5 asks interest in the user's future to estimate motivation, which implies mental health condition. The scenario content has been changed slightly depending on the weather and time of day. We also conducted a

Table 3. Scenario script of the practice and the type of the voice-calls (User study 3).

Turn no.	Type	Voice-call
1	S	“Hi. I’m Bono-chan. Nice to meet you. What is the date today?”
2	OQ	“Thank you very much. So what are your hobbies?”
3	OQ-R	“I see. I want to do that too.”
4	OQ	“Ok. Then what is your favorite color?”
5	OQ-R	“I see. My favorite color is ultramarine blue.”
6	S	“It’s about time. So let’s end this conversation now.”

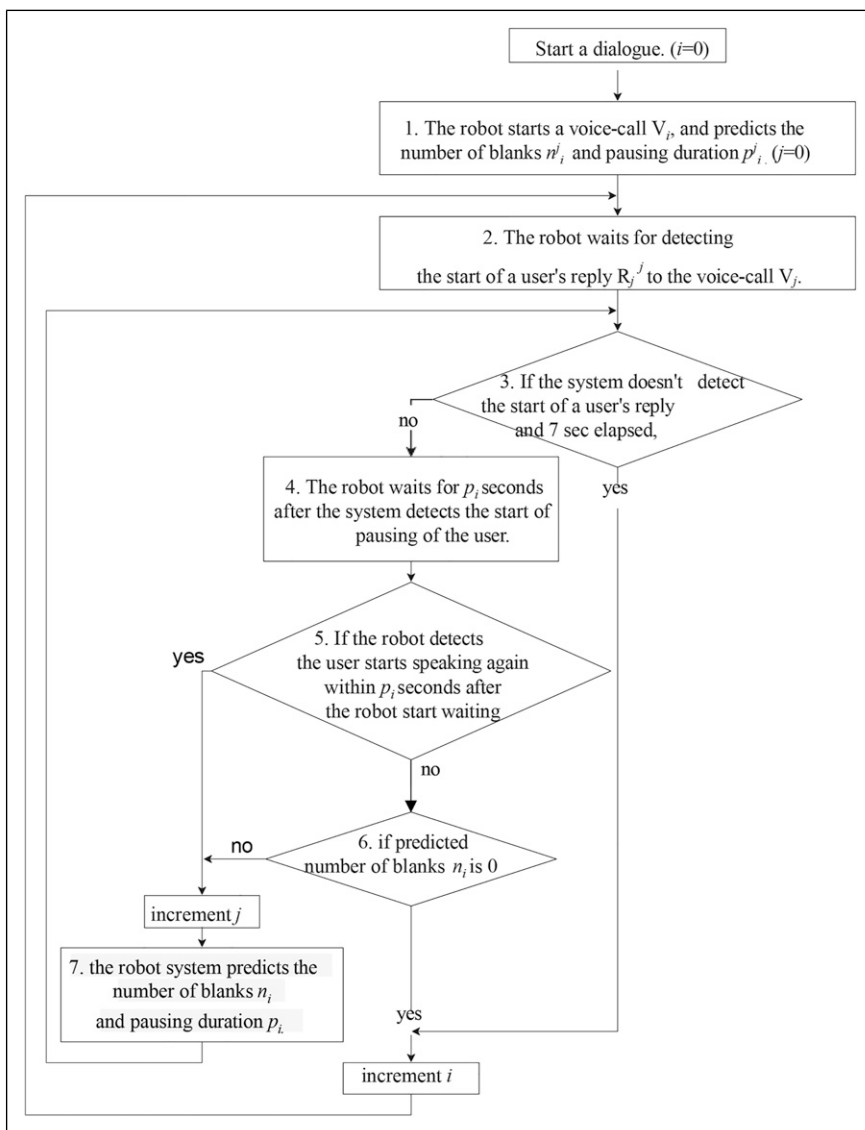


Figure 16. The flowchart of User Study 3.

questionnaire after the dialogue under each condition of controlling the robot.

Each participant had a dialogue a day and the user study for each participant was conducted once a day for a total of

2 days. On the final day, an interview was conducted, and participants answered their impressions of the content of the dialogue with the robot and the timing of start a voice-call of the robot etc. The procedure of the user study was as follows.

Table 4. Example of scenarios in User Study 3.

Turn no.	Type	Topic	Voice-call
1	S	1: Asking health status	“Good morning. Nice to meet you.”
2	OQ	1	“How are you feeling now?”
3	OQ-R	1	“I see. Thank you for sharing.”
4	S	2: Invite for a walk	“It’s nice weather today.”
5	S-R	2	“I feel like go for a walk.”
6	OQ	2	“Where do you want to go for a walk?”
7	OQ-R	2	“I see. I will go to a park near my house.”
8	S	3: Asking about yesterday	“Yes. By the way, I played soccer with my friend yesterday.”
9	S-R	3	“It was so cool.”
10	OQ	3	“Tell me a fun thing you did yesterday.”
11	OQ-R	3	“I see. That’s nice.”
12	S	4: Asking about meal	“It’s almost lunchtime.”
13	S-R	4	“I want to eat warm food for lunch.”
14	OQ	4	“What do you want to eat for lunch?”
15	OQ-R	4	“OK, I want to eat that with you.”
16	S	5: Asking about next year	“By the way, 2021 is already October.”
17	S-R	5	“I would like to be a robot that can play an active role next year.”
18	OQ	5	“OK. So, what are your goals for next year?”
19	OQ-R	5	“I see. Yes, let’s have a good year with each other.”
20	S	6: Closing	“It’s about time to end. Let’s close this dialogue now.”

- 1st day

1. The practice of a dialogue with the robot.
2. First dialogue with the individualized turn-taking system*.
3. A participant fills in the questionnaire to evaluate the dialogue.

- 2nd day

1. The practice of a dialogue with the robot.
2. Second dialogue with the not individualized turn-taking system*.
3. The participant fills in the questionnaire to evaluate the dialogue
4. Interviews about the dialogues with the robot.

The order of the two dialogues (individualized/not individualized) was randomized in consideration of counterbalance.

In this study, 10 older adults (from 60 to 80 years old) of four men and six women, and 10 non-elderly adults (younger than 60 years old) and five men and five women participated. Eight of the older adults were participants in the second user study. The users were instructed to answer to the robot’s voice-call in as much detail as possible before starting the user study.

Result: Figure 17 summarizes the number of overlaps during the dialogue. We counted the number of times the

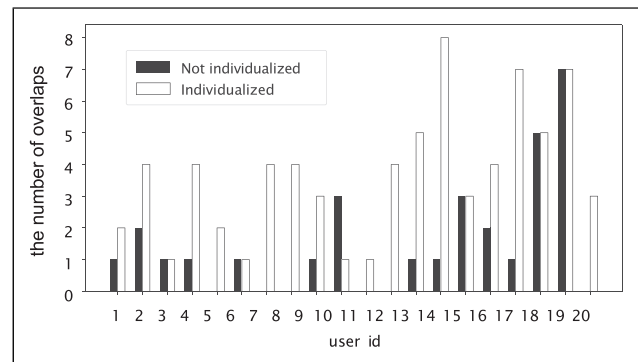


Figure 17. The number of overlaps for each methods (individualized method vs. generalized method) in User Study 3.

robot’s voice-call and the user’s speech overlapped when the robot started its voice-calls. First, under the condition that the method of individualizing was not adopted, the robot’s voice-calls and user’s reply overlapped 1.5 times on average. By contrast, in the case of the dialogue system, which employed the method of individualizing, the number of overlaps was 3.8 times, on average.

Next, we describe the prediction results of the blank duration and the number of blanks during the user’s speech. We calculated the difference between the predicted value and the observed value (we defined the difference as the prediction error). To ensure the change of the prediction error as a result of online learning, we compared the prediction error at the beginning and at the end of one dialogue

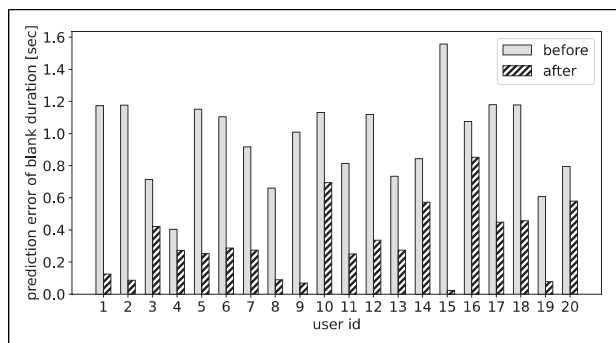


Figure 18. The prediction error of the duration of blank in the case of replies to non-open questions (predicted in the early stage vs. predicted in the late stage).

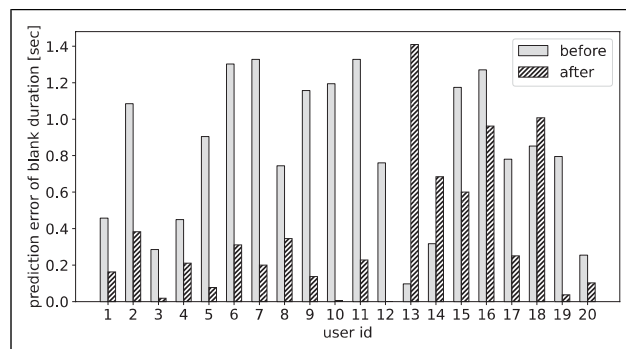


Figure 20. The prediction error of the duration of blank in the case of replies to open questions (predicted in the early stage vs. predicted in the late stage).

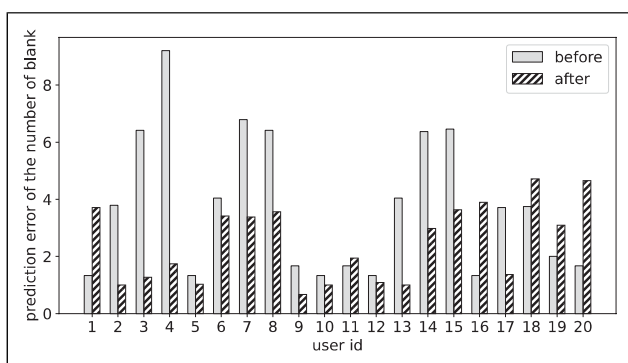


Figure 19. The prediction error of the number of blanks in the case of replies to non-open questions (predicted in the early stage vs. predicted in the late stage).

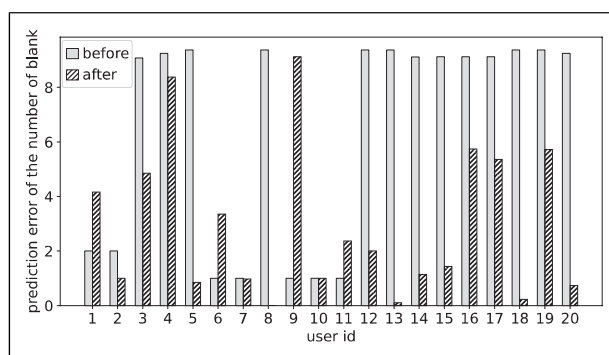


Figure 21. The prediction error of the number of blanks in the case of replies to open questions (predicted in the early stage vs. predicted in the late stage).

user study consisting of 20 voice-calls. The list of the voice-calls is shown in Table 4. Specifically, the beginning stage is a set of 4 turns (turn no 4–7) of a scene about the weather and a walk, except for the first greeting and confirmation of physical condition. Also, the end stage is the set of 4 turns (turn no 16–19) of a scene about the resolution of this year or next year, except for the greeting at the end of the dialogue. Each set includes three voice-calls of not open-question and one voice-calls with open question.

Figures 18–23 shows a graph comparing the prediction errors of the blank duration and the number of blanks at the beginning and at the end of the user study.

Users 1 to 10 were younger than 60 years old, and users 11 to 20 were older than 60 years old. In the case of the blank duration and the number of blanks included in the user's reply to the robot's voice-call of not open-question, the prediction error of the blank duration decreased in the end stage compared to the beginning (Figures 18 and 19). By contrast, regarding the number of blanks, there were some users whose prediction errors were large in the latter

half of the voice-call. Second, in the case of the robot's voice-call with open question, about half of the participants' prediction error of the blank time did not change from the beginning stage. For some participants, the prediction error of the number of blanks was smaller than at the beginning; however, the maximum value of the prediction error was about six (Figures 20 and 21).

The box plot in Figures 22 and 23 show the results of each user's reactions: duration and number of blanks for OQ and NOQ. These figures also include the values predicted at the beginning and at the end of the user study. In some user cases, the plots of the value predicted at the end of the user study were closer to the observation results than the plots of the value predicted at the beginning. In the case of the prediction of pause duration, the values predicted at the beginning were larger than those predicted at the end of the user study. The figure suggests that after the system learned each user's reactions in this user study, the predicted value tended to be shorter than the observed values. This result may indicate that some users feel the robot does not listen to the user. Therefore, we will consider the method to cope

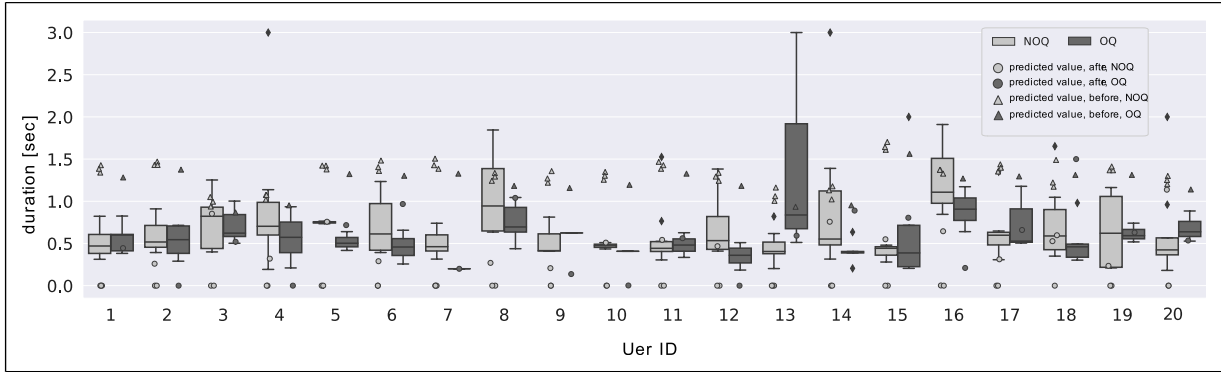


Figure 22. The observed values of duration of blank according to the type of robot’s voice-call and the values predicted by the robot system at the beginning (before) and at the end of the user study (after) for each user.

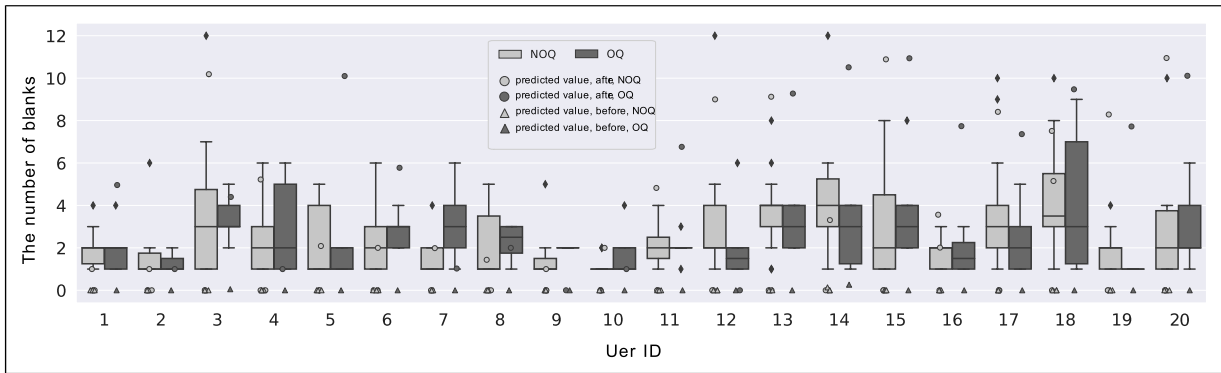


Figure 23. The observed values of the number of blanks according to the type of robot’s voice-call and the values predicted by the robot system at the beginning (before) and at the end of the user study (after) for each user.

with this issue, such as waiting longer than predicted before starting a voice-call in future work.

Figure 24 shows the score of the degree to which the user tried to adjust his/her speaking pace to the robot. There was no difference between the two conditions: the individualized condition and the generalized condition, although a few more participants tried to adjust the timing to the robot in the individualized condition. Figure 25 shows the relationship between the number of overlaps and the score of willingness to continue dialogue. Overall, 6 out of 20 participants overlapped with the robot more than five times, and the evaluations of those six participants were almost the same as the scores of other participants.

The scenario used in User Study 3 included a voice-call to check a health state of a user. Table 5 shows the responses

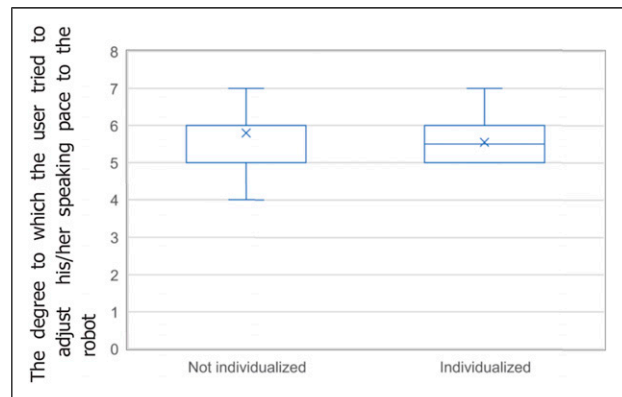


Figure 24. The score of the degree to which the user tried to adjust his/her speaking pace to the robot in User Study 3.

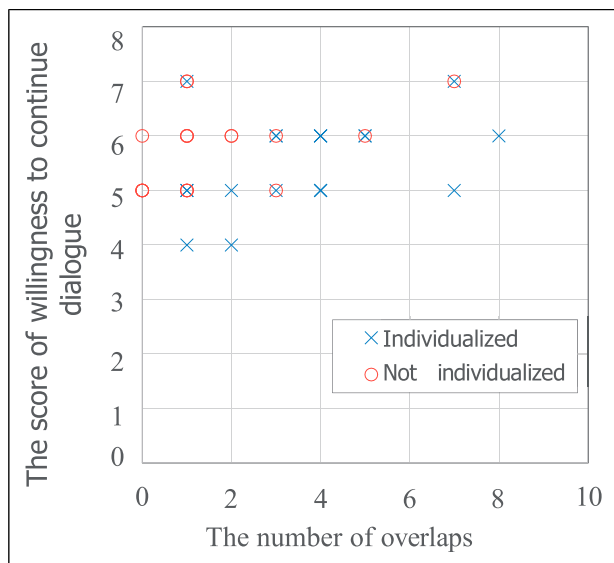


Figure 25. The score of willingness to continue dialogue versus the number of overlaps.

of all users to the robot's voice-call. The content of the reply was transcribed after the user study. We confirmed that all users were able to respond to the voice-call of the robot in this user study. In the future, we will build a system which allows to realize such the health status check. We will build a robot dialogue system (local) set at home and a system that estimates the health condition based on the data sent from the local robot system to a remote server and decides a voice-call to check the physical condition according to the individual user.

Discussions

Individual differences in responses to a scenario-based dialogue robot in User Study 1 and 2

Some characteristic reactions that are thought to be caused by user's individual differences were observed in this user study. First, some user sometimes took a longer time to start talking after the robot's voice-call. Next, another user gave more back-channel feedback (e.g. saying "yeah" during the robot's voice-call). Back-channeling is a sign of the attitude of the listener to the speech, and the manner in which back-channeling is given is thought to differ according to the character of individuals: nodding, facial expression, or saying something. The user's back-channel was distinguished from the user's interrupted speech because the user vocally gave his/her back-channels, which repeatedly caused the robot's voice-call to stop and made the user confused. Next, based on the content of the users' speech, which were transcribed after the user study, a user answered the robot's question and then returned the same question to

the robot more times (ex. Saying, "How about you? Bono-chan."). The amount of talking of the user's speech was higher than that of other users. We prepared some scenarios to handle the case in which a user returns the question. For example, the robot says, "My favorite color is ultramarine," in response to the user's answer to the robot's previous voice-call, "What is your favorite color?" (Table 6). However, a repeated lack of response from the robot to the question of the user lowers the user's satisfaction level, as they feel their question cannot be answered.

User's reaction to match the timing of the robot's voice-call in User Study 3

In the case of using the individualized method, some participants felt it was difficult to adjust the timing of the speech to the robot's side because the robot also tried to adjust its timing. For example, some answers obtained during the interview were: (A) "I felt that the robot was waiting for me when my speech was short. But I did not know whether it was better to talk longer or shorter, so I could not adjust the timing well," (B) "It was easier to talk if the timing of voice-call was constant because I can read the timing of the robot." However, according to the results of the questionnaire, most of the participants answered that they adjusted their pace of speaking to the robot.

This result of Figure 24 and the interview may indicate that many of the participants in this user study were accustomed to interacting with robots. In future studies, we will include users whose attributes are different from the participants of this user study, confirm the user's attributes in advance, and investigate their reaction while comparing their attributes and their subjective evaluation of the robot.

In the condition of the generalized method, the average number of times the robot's voice-call and the user's speech overlapped was about 1.5 times (there were 20 voice-calls in one user study). In the interview about the system with the generalized condition, there were few negative references to the overlap. Thus, it is probable that the turn-taking was successful in the dialogue of this condition. When the robot's voice-call and a user's speech overlap, the robot's voice-call "Please go ahead" allowed users to know the situation in which they overlapped with each other, and the robot gave the turn to the user.

Comparison of the results of User Study 1, 2, and 3

One of the causes of the increased overlap in User Study 3 was a user's attempt to break a silence. In User Studies 1 and 2, the robot and a user overlapped because the user tried to break a silence. Similarly, some user's speech overlapped with the robot's voice-call because the robot waited for 2.0 s after the user's speech. Such a situation was

Table 5. User's answer to the voice-call about health.

User id	today's feeling
1	I feel good today. It's normal but good.
2	I'm great.
3	Yes, I feel very good today. I was especially looking forward to talking with Bono-chan, so I'm very happy to see your cute face.
4	I feel happy today.
5	I'm fine as usual today.
6	I slept well yesterday so I feel good today.
7	I don't feel very good today.
8	I feel a little sleepy. Other than that, I feel refreshed.
9	I'm feeling good today. I had a good sleep last night, so I quite active today.
10	I'm fine.
11	Yes. It's very good today.
12	As usual, I have a very energetic and nice day.
13	I'm feeling better than I expected.
14	The weather is nice so I'm fine. It feels good and feels very good.
15	It's pretty good. The weather is nice and it feels good. The scent of osmanthus still remained, which made me feel a little happy.
16	It's a refreshing day in autumn.
17	I'm fine. I spend my time thinking about various things.
18	I feel so good today. But meeting Bono-chan made me feel really good.
19	The weather is nice today so I feel very good.
20	Well, it's normal.

Table 6. Example of the scenario to handle the User's repeating the same question to the robot (User Study 2).

Speaker	(Assumed/predetermined) content of the speech
Robot (Bono)	What is your favorite color?
User	My favorite color is blue. How about you, Bono?
Robot (Bono)	My favorite color is ultramarine.

observed under the conditions of the generalized method. By contrast, under individualized conditions, the robot learned that the blank duration was shorter than predicted and tried to shorten the waiting duration after the end of the user's speech, which resulted in the overlap between the user and the robot. Some participants felt that the robot limited the time required for the user's speech. We assume that overlapping with a robot does not necessarily reduce the satisfaction of the dialogue based on Figure 25. Interview responses showed that several participants felt that the robot's voice-call, "Please go ahead," was as if the robot was saying "speak more."

Key contribution of this paper

We built a scenario-based dialogue system in which the robot's voice-call was preset while the turn-taking was automated. Then, we conducted user studies of dialogues between the robot and older adults using the turn-taking system and found that dialogues including multiple turn-taking (up to 20 turns), can be successfully performed with the system.

We observed users' reactions to the robot's voice-call (i.e. the time before beginning to reply to the robot, the frequency of the back-channeling, and returning the same questions to the robot) differed across individuals. Moreover, other individual differences of users' behavior in relation to the robot were observed when their speeches overlapped with the robot and when silence occurred during dialogue. In User Study 3, we used the dialogue system, which was updated so that the system learns the pauses of the user during dialogue and controls the start time of voice-calls to the user based on the result of this learning. The system estimates the length and the number of pauses based on the content of voice-calls. We observed that some users had conversations at a good tempo, while for other users, the number of overlaps between the robot and the user increased. We thought these differences were a result of the users' attitude, that is, whether they tried to adjust to the robot or not. As a contribution, our paper found that some participants tried to adjust their talking pace to the robot, while some others did not. We also found that there was a lack of information regarding the appropriate estimation of the pausing segments of each individual. The

individual differences in reactions to the robot are considered to be one of the keys to estimate appropriate pausing segments. It is important that experimental research clarify the information needed for estimating appropriate pausing segments as the next step of the research.

Application and limitation

Of the five technologies described in the Chapter 2, we studied a method to keep conversations between older adults and the robot going for longer period of time as our first step. This chapter describes what we found and limitations. Furthermore, we observed user responses related to the next steps so we discuss future works.

- (i) **Method to keep conversation between the older adult and a robot for longer period of time:** We constructed the scenario-based dialogue system with an automated turn-taking function and confirmed that the robot and older users tended to keep a dialogue of 17–20 turns, which is longer than our previous study. In User Study 3, we implemented a system of predicting users' pause duration within the dialogue system. For some users, the system predicted the user's pause duration and maintained a good dialogue tempo according to the interviews. We also found that some users tried to match their speech timing to that of the robot's, causing the voice-call to overlap with the user's reply. As a limitation, three User Study were conducted remotely due to COVID-19 protocols. There are differences in persuasiveness between a real robot and a virtual robot; in this study we conducted daily conversations, and there may be differences in the responses of users depending on the content of the dialogue. To carry out the dialogue user studies using video chat software, older adults who were accustomed to digital devices such as smartphones and laptops participated in the user studies. Talking to robots was not a difficult scenario for them, and users did not seem to hesitate when talking to the robot. We will proceed with the plan to carry out the user study at home.
- (ii) **Method to prolong the older adult's interests and to maintain engagement with the robot:** Scenario-based dialogue can be continued so long as the content is prepared and the user's interest lasts. Based on the interviews and experiments conducted in this experiment, we found the following issues related to long-term adherence: some users tried to continue the dialogue even after overlaps. However, the inappropriate timing of responses and poor back-channeling are thought to decrease

users' interests. We will conduct a detailed study to determine how to prolong the user's interest in the robot as the next step. In line with this goal, we are considering a method of adding scenarios associated with the season and the user's hobbies.

- (iii) **Towards voice-call based on health status of older users:** In User Study 3, all the participants answered to the robot's voice-call asking about the user's health status. However, in dialogues with caregivers or families at home, older users tended to respond to questions about their health condition or problems. We will verify a method of indirectly asking about the health condition of older adults. As another project of a dialogue robot system, we are developing a speech recognition-based dialogue system.⁴² We are considering integrating the speech recognition-based system into our scenario-based dialogue system to deal with the specific situation which needs to use speech recognition to realize the appropriate robot's voice-call to the older user.

Integration of sensor-based AAL and scenario-based dialogue system

The developed system is designed for monitoring health conditions of older adults. It would be integrated to sensor-based AAL system. We have developed sensor-driven scenario-based dialogue system.¹⁸ The mat sensors are located on the bed and based on the state transition of the user on the bed, the robot talks to the user on the bed using scenario-based dialogue system. Also, the scenario-based dialogue system could be implemented where one scenario is selected among multiple candidate scenarios based on the output of the sensors. If the temperature sensor detects that the temperature is hot, the scenario which begins with "Today is hot, isn't it" may be selected.

Conclusion

In this study, our objective was to realize a robot dialogue system that supports the independent living of older adults. We plan to use the voice-calls of a robot to encourage older adults to speak and estimate the health condition of the person based on the speech. In this way, it is most important to continue the dialogue for longer periods, since more information might be extracted if conversations last longer. We have studied the method to prolong the conversation by managing the rhythm of turn-taking. The proposal and verification results of the turn-taking method for the scenario-based dialogue have been reported in this paper. We built a scenario-based dialogue system in which the robot's voice-call was preset while the turn-taking system

was automated. The robot system monitored the sound level acquired by the microphone during the user's speech and detected pauses in the user's speech. The robot's waiting time after detecting the pause was determined based on the prediction of the pause duration and the number of pauses.

We conducted three user studies in which the user and robot had a scenario-based dialogue. First, we collected the pause data of the users to obtain training data to train the pause predictor in User Study 1. In User Study 2, we built a prototype of the system of scenario-based dialogue, and the responses of the users were observed. In User Study 3, we used a pause predictor to predict the user's pauses, and the responses of the users were compared with the dialogue system without the individual pause predictor.

The content of the scenario was evaluated as natural in the studies. Regarding turn-taking in the dialogue with the robot, users' speech and the robot's voice-call sometimes overlapped in some users. We collected data on pause duration, the number of pauses, the speech duration, and the word count of the users' speech, and found that there was a difference in these features according to the types of the robot's voice-call. Individual differences were also found. Based on these results, we added the function of predicting an individual's pause duration and the number of pauses to the dialogue system as an approach to smoother turn-taking. With the updated scenario-based control method, the number of overlap was 1.5 times on average. In the cases where the individualized method is employed, however, the number of overlaps was higher. Some of the users who overlapped with the robot's speech tried to adjust the timing of starting their own speech by predicting the robot's waiting time and failed.

Further consideration is needed to yield any findings regarding individual differences in responses to the robot's voice-call. Additionally, we are planning to integrate the voice-call system with a system that estimates the health condition of older adults, as well as to build a system with a scenario determination method according to the individual's condition.

Acknowledgements

We thank Fonobono Research Institute for help with recruiting participants, enabling us to conduct our user study smoothly. We are also grateful for all the participants and staff for this study.

Author contributions

NM and MO-M researched literature and conceived the study. ST, NM, KT, and MO-M developed of the system and the robot. KK, ST, NM, and KT updated the system design and the program. NM and MO-M gained ethical approval. KK, ST, NM, and KT conducted user study, data analysis and data collection. KK and NM wrote the first draft of the manuscript. All authors reviewed and

edited the manuscript and approved the final version of the manuscript.

Declaration of conflicting interests

The author(s) declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was supported by the JSPS KAKENHI (Grant Numbers JP18KT0035, JP19H01138, JP20H05022, JP20H05574, JP20K19471, JP22H00544, JP22H04872) and the Japan Science and Technology Agency (Grant Numbers JPMJCR20G1, JPMJST2168, JPMJPF2101, JPMJMS2237).

Guarantor

MO-M

ORCID iD

Kazumi Kumagai  <https://orcid.org/0000-0003-3191-3219>

References

1. Anghel I, Cioara T, Moldovan D, et al. Smart environments and social robots for age-friendly integrated care services. *Int J Environ Res Public Health* 2020; 17: 3801.
2. Ghayvat H, Mukhopadhyay S, Shenjie B, et al. Smart home based ambient assisted living: Recognition of anomaly in the activity of daily living for an elderly living alone. In *IEEE International Instrumentation and Measurement Technology Conference (I2MTC)*. IEEE, 2018.
3. Yuan F, Klavon E, Liu Z, et al. A systematic review of robotic rehabilitation for cognitive training. *Front Robotics AI* 2021; 8: 105.
4. Ciciirelli G, Marani R, Petitti A, et al. Ambient assisted living: A review of technologies, methodologies and future perspectives for healthy aging of population. *Sensors* 2021; 21: 3549.
5. Begum M, Wang R, Huq R, et al. Performance of daily activities by older adults with dementia: The role of an assistive robot. In *IEEE 13th International Conference on Rehabilitation Robotics (ICORR)*. IEEE, 2013.
6. Heerink M, Kröse B, Evers V, et al. *The Influence of Social Presence on Acceptance of a Companion Robot by Older People* 2008.
7. Obayashi K and Masuyama S. *Pilot and Feasibility Study on Elderly Support Services Using Communicative Robots and Monitoring Sensors Integrated with Cloud Robotics*. *Clinical Therapeutics* 2020.

8. Al-Khafajiy M, Baker T, Chalmers C, et al. Remote health monitoring of elderly through wearable sensors. *Multimedia Tools Appl* 2019; 78(17): 24681–24706.
9. Li J, Ma Q, Chan AH, et al. Health monitoring through wearable technologies for older adults: Smart wearables acceptance model. *Appl Ergon* 2019; 75: 162–169.
10. Maskeliūnas R, Damaševičius R and Segal S. A review of internet of things technologies for ambient assisted living environments. *Future Internet* 2019; 11(12): 259.
11. Rodrigues MJ, Postolache O and Cercas F. Physiological and behavior monitoring systems for smart healthcare environments: A review. *Sensors* 2020; 20: 2186.
12. Malwade S, Abdul SS, Uddin M, et al. Mobile and wearable technologies in healthcare for the ageing population. *Comp Methods Programs Biomedicine* 2018; 161: 233–237.
13. Bajones M, Fischinger D, Weiss A, et al. Hobbit: Providing fall detection and prevention for the elderly in the real world. *J Robotics* 2018; 2018: 1–20.
14. Gasteiger N, Ahn HS, Fok C, et al. Older adults' experiences and perceptions of living with bomy, an assistive daily care robot: a qualitative study. *Assistive Technol* 2021; 34: 1–11.
15. Graf B, Reiser U, Hägele M, et al. Robotic home assistant careo-bot[®] 3-product vision and innovation platform. In IEEE Workshop on Advanced Robotics and its Social Impacts. IEEE, 2009.
16. Shrivastava R and Pandey M. Real time fall detection in fog computing scenario. *Cluster Computing* 2020; 23(4): 2861–2870.
17. Miyake N, Kumagai K, Tokunaga S, et al. Towards practical use of bedside sensing/voice-calling system for preventing falls. In International Conference on Human-Computer Interaction. Springer.
18. Miyake N, Shibukawa S, Masaki H, et al. User-oriented design of active monitoring bedside agent for older adults to prevent falls. *J Intell Robot Syst* 2020; 98(1): 71–84.
19. Tani M, Hirayama A, Torimoto K, et al. Guidance on water intake effectively improves urinary frequency in patients with nocturia. *Int J Urol* 2014; 21(6): 595–600.
20. Lazarus RS. *From Psychological Stress to the Emotions: A History of Changing Outlooks*. 1993.
21. Bone D, Lee CC and Narayanan S. Robust unsupervised arousal rating: A rule-based framework with-knowledgeinspired vocal features. *IEEE Transact Affect Comp* 2014; 5(2): 201–213.
22. Schmidt J, Janse E and Scharenborg O. Perception of emotion in conversational speech by younger and older listeners. *Front Psychol* 2016; 7: 781.
23. Cannizzaro M, Harel B, Reilly N, et al. Voice acoustical measurement of the severity of major depression. *Brain Cognit* 2004; 56(1): 30–35.
24. Moore E, Clements M, Peifer J, et al. Analysis of prosodic variation in speech for clinical depression. In Proceedings the 25th Annual International Conference the IEEE Engineering Medicine Biol Society. IEEE.
25. Mundt JC, Snyder PJ, Cannizzaro MS, et al. Voice acoustic measures of depression severity and treatment response collected via interactive voice response (ivr) technology. *J Neurolinguistics* 2007; 20(1): 50–64.
26. Vicsi K, Sztah'ó D and Kiss G. Examination of the sensitivity of acoustic-phonetic parameters of speech to depression. In IEEE 3rd International Conference on Cognitive Infocommunications (CogInfoCom). IEEE, 2012.
27. Wang J, Zhang L, Liu T, et al. Acoustic differences between healthy and depressed people: a cross-situation study. *BMC Psychiatry* 2019; 19(1): 1–12.
28. Yang Y, Fairbairn C and Cohn JF. Detecting depression severity from vocal prosody. *IEEE Transact Affect Comput* 2012; 4(2): 142–150.
29. Fanshel S and Bush JW. A health-status index and its application to health-services outcomes. *Operations Research* 1970; 18(6): 1021–1066.
30. Razuri JG, Sundgren D, Rahmani R, et al. Speech emotion recognition in emotional feedback for human-robot interaction. *Int J Adv Res Artif Intelligence (IJARAI)* 2015; 4(2): 20–27.
31. Ruvolo P, Fasel I and Movellan J. Auditory mood detection for social and educational robots. In IEEE International Conference on Robotics and Automation. IEEE, 2008.
32. Zeng Z, Pantic M, Roisman GI, et al. A survey of affect recognition methods: Audio, visual, and spontaneous expressions. *Pattern Analysis and Machine Intelligence, IEEE Trans* 2009; 31(1): 39–58.
33. Gruebler A, Berenz V and Suzuki K. Coaching robot behavior using continuous physiological affective feedback. In 11th IEEE-RAS International Conference on Humanoid Robots. IEEE, 2011.
34. Zarakı A, Khamassi M, Wood LJ, et al. A novel reinforcementbased paradigm for children to teach the humanoid kaspar robot. *Int J Soc Robotics* 2019; 2019: 1–12.
35. Li Y, Ishi CT, Inoue K, et al. Expressing reactive emotion based on multimodal emotion recognition for natural conversation in human–robot interaction. *Adv Robotics* 2019; 33(20): 1030–1041.
36. Gordon G, Spaulding S, Westlund JK, et al. *Affective Personalization of a Social Robot Tutor for Children's Second Language Skills*. AAAI.
37. Kumagai K, Baek J and Mizuuchi I. A situation-aware action selection based on individual's preference using emotion estimation. In Proceedings of the 2014 IEEE International Conference on Robotics and Biomimetics.
38. Shinohara S, Omiya Y, Hagiwara N, et al. Case studies of utilization of the mind monitoring system (mimosys) using voice and its future prospects. *ESMSJ* 2017; 7(1): 7–12.
39. Sacks H, Schegloff EA and Jefferson G. A simplest systematics for the organization of turn taking for conversation.

- In Studies in the organization of conversational interaction. Elsevier, 1978.
40. Kawasaki M, Yamada Y, Ushiku Y, et al. Inter-brain synchronization during coordination of speech rhythm in human-to-human social interaction. *Sci Rep* 2013; 3(1): 1–8.
 41. Tokunaga S and Otake-Matsuura M. Development of a dialogue robot bono-06 for cognitive training of older adults. *Gerontechnology* 2020; 19: 1.
 42. Tokunaga S, Tamura K and Otake-Matsuura M. A dialogue-based system with photo and storytelling for older adults: Toward daily cognitive training. *Front Robotics AI* 2021; 8: 644964–645015.
 43. ReadSpeaker Holding BV. *ReadSpeaker Web API*. <https://www.readspeaker.com/>
 44. Likert R. *A Technique for the Measurement of Attitudes*. Archives of Psychology 1932.
 45. Masumura R, Asami T, Masataki H, et al. Online end-of-turn detection from speech based on stacked time-asynchronous sequential networks. *Interspeech* 2017; 2017: 1661–1665.
 46. Levinson SC and Torreira F. Timing in turn-taking and its implications for processing models of language. *Front Psychol* 2015; 6: 731.
 47. Liu C, Toshinori CI and Ishiguro H. Turn-taking estimation model based on joint embedding of lexical and prosodic contents. In *Interspeech*.
 48. Ishimoto Y, Teraoka T and Enomoto M. End-of-utterance prediction by prosodic features and phrase-dependency structure in spontaneous japanese speech. In *Interspeech*.
 49. Hara K, Inoue K, Takanashi K, et al. Prediction of turn-taking using multitask learning with prediction of backchannels and fillers. *Listener* 2018; 162: 364.
 50. Skantze G. Towards a general, continuous model of turn-taking in spoken dialogue using lstm recurrent neural networks. In Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue. pp. 220–230.
 51. Roddy M, Skantze G and Harte N. Multimodal continuous turn-taking prediction using multiscale RNNs. In Proceedings of the 20th ACM International Conference on Multimodal Interaction. pp. 186–190.
 52. Welch BL. The significance of the difference between two means when the population variances are unequal. *Biometrika* 1938; 29(3/4): 350–362.
 53. Vapnik V. Pattern recognition using generalized portrait method. *Automation Remote Control* 1963; 24: 774–780.