Paired Conversational Agents for Easy-to-Understand Instruction

Wataru Akahori
NTT Service Evolution Labs.
wataru.akahori.cd@hco.ntt.co.jp

Hiroaki Sugiyama
NTT Communication Science Labs.
hiroaki.sugiyama.kf@hco.ntt.co.jp

Hiroya Minami
NTT Service Evolution Labs.
hiroya.minami.gs@hco.ntt.co.jp

Asuka Miyake
NTT Service Evolution Labs.
asuka.miyake.zw@hco.ntt.co.jp

Masahiro Watanabe
NTT Service Evolution Labs.
masahiro.watanabe.xd@hco.ntt.co.jp

ABSTRACT

Conversational agents such as those hosted by the 'smart speakers' have become popular over the last few years. Although users can accomplish tasks as if they were asking a person, users still have problems in utilizing conversational agents effectively. To address this problem, some proposals explain how to input agent requests by using visual information such as instruction manuals and displays. However, such instructions create problems such as occupying the hands and eyes. The purpose of this study is to effectively enhance request entry by issuing instructions for use in an easy-to-understand manner without using visual information. Our proposal uses a pair of conversational agents, one is called the main agent, and the other is called the sub-agent, that have different voice types. Experiments show that agent pairing yields easier to understand instructions than using just

CHI'19 Extended Abstracts, May 4–9, 2019, Glasgow, Scotland UK

ACM ISBN 978-1-4503-5971-9/19/05.
https://doi.org/10.1145/3290607.3312794

the main agent. Furthermore, experiments also show that use instructions are easier to understand if the sub-agent reads aloud specific examples of use.

KEYWORDS

conversational agent; paired agents; instruction; smart speaker.

INTRODUCTION

Conversational agents have become pervasive in everyday life. For example, the smart speakers such as the Amazon Echo and Google Home have become widespread over the last few years. Several HCI studies have reported that although users made requests to the conversational agent as if they were asking a person, the agent responded in manner different from what was expected. For example, the conversational analysis of Porcheron *et al.* [7] indicated that since the response from the system did not consistently follow any user's input, the system-person dialogue was fundamentally different from the person-person dialogue. Furthermore, Luger and Sellen [5] found that those with lower levels of technical knowledge had a high level of frustration because their high expectations of the conversational agents created mismatch in the perception of system intelligence.

To address this problem, many approaches explain how to make requests by using visual information such as instruction manuals and displays. However, they have disadvantages such as occupying the hands and eyes. Although audio-only information presentation is required to overcome the disadvantages posed by the visual information, due to the working-memory limitation [6], the explanation provided only with audio is hardly understood by users compared with both audio and visual. On the other hand, several studies have reported that using multiple agents to present information had a positive impact on understanding the information [1, 4, 8]. However, these studies used screen agents as visual information, so they did not address the effectiveness of multiple talking agents that are heard but not seen.

In this paper, we introduce a presentation method that makes it easy to understand instructions without using visual information. Based on the previous studies [1, 4, 8], we propose paired invisible conversational agents, one is called the main agent, and the other is called the sub-agent, that have different voice types. We conducted a Wizard-of-Oz experiment to verify the effectiveness of the proposed method in terms of instruction understandability and cognitive load requirements.

Our main contributions include:

- Verifying the effectiveness of audio-only instruction in terms of understandability by introducing a sub-agent who has a different role from that of the main agent.
- Expanding the application range of audio-only information presentation from smart speakers.

Table 1: An example of a self-introduction and an explanation.

Type of speech	Reading aloud	Q&A
Role of sub-agent	Reading aloud specific example	Asking main agent
Self-introduction	**Main:** The next function will be explained by two people. **Sub:** I will give you a specific example of how to use the function.	**Main:** The next function will be explained by two people. **Sub:** I will ask about how to use the function. **Main:** I will answer the queries.
Explanation	**Main:** I will explain how to use the weather forecast function. With this function, you can search the weather forecast for up to 8 days ahead by saying the date and location. Specifically, if you say **Sub:** Tell me the weather in Chiyoda-ku, Tokyo today. **Main:** You will access this function. You can check the weather of the area by setting the zip code of your area with the application on the smart phone. It ends with the description just given.	**Main:** I will explain how to use the weather forecast function. With this function, you can search the weather forecast for up to 8 days ahead by saying the date and location. **Sub:** Interesting, how do I use it? **Main:** Specifically, if you say "Tell me the weather in Chiyoda- ku, Tokyo today," you will access this function. **Sub:** What else can I do? **Main:** You can check the weather of the area by setting the zip code of your area with the application on the smart phone. It ends with the description just given.

INSTRUCTION PRESENTATION METHOD

To make instructions easier to understand, we use paired conversational agents. Several studies have reported that presenting information from multiple agents had a positive impact on understanding the information. For example, it may be easier for the listener to separate the types of information if the two agents speak alternately [1]. Furthermore, the conversational representation may make sentence construction simpler and easier to understand [4]. Thus, the proposed method uses paired agents that use two types of speech to present instructions. The main agent provides instructions and the sub-agent supports the main agent.

Table 1 shows an example of a self-introduction and an explanation. First, the proposed method starts with self-introduction of the two agents in order to reduce user's cognitive loads. Second, the proposed method assigns two kinds of roles for the sub-agent: reading aloud a specific example and posing questions to the main agent. The first role aims to make it easier to focus on the voice command which is necessary to make a request to the conversational agent. In the example of Table 1, the sub-agent issues the command, "Tell me the weather in Chiyoda-ku, Tokyo today." The other role is based on research that confirms that understanding is enhanced by using concrete questions to

Table 2: Presentation styles examined.

Style	1	2	3	4	5	6	7	8
Self-introduction	without	without	without	without	with	with	with	with
Number of agents	one	two	one	two	one	two	one	two
Type of speech	reading aloud	reading aloud	Q&A	Q&A	reading aloud	reading aloud	Q&A	Q&A

present context information [4]. In the example of Table 1, the sub-agent asks the main agent, "How do I use it?" The sub-agent also asks, "What else can I do?"

EXPERIMENT

We conducted a Wizard-of-Oz experiment to verify the effectiveness of the proposed method in terms of understandability of instructions and cognitive loads. A three-factor mixed ANOVA test was performed with self-introduction (with or without) as the between-subjects factor and the number of agents (one or two) as the within-subjects factor and type of speech (reading aloud or Q&A) as another within-subjects factor. Thus, self-introduction and explanation were presented using the $2 \times 2 \times 2 = 8$ styles shown in Table 2.

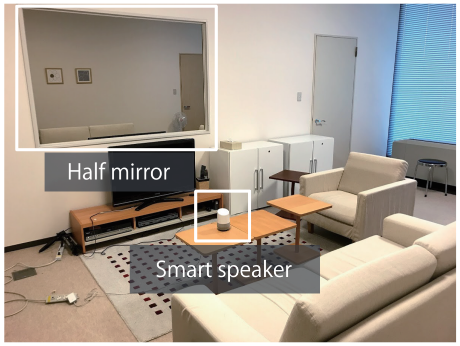

Figure 1: Experimental setup.

Methods

18 Japanese participants (nine males and nine females, ranging from 21 to 56 years old) who have never used a smart speaker were recruited. The participants were divided into two groups. The group without self-introduction consisted of nine participants (five males and four females) and the group with self-introduction consisted of the remaining nine participants (four males and five females).

The experimental setup is shown in Figure 1. The speech was created using NTT's speech synthesis software [3]. The main agent used a female voice and the sub-agent used a male voice. We used Google Home as the smart speaker. The PC and smart speaker were connected by Bluetooth. Agent utterances were generated from audio files in the PC. The experimenter who controlled the utterances sat outside the room behind the half mirror so as to hide the operations from the participants.

Participants were asked to use the function as instructed by the smart speaker. The experimental procedure was as follows. First, for the group with self-introduction, each participant was presented with the self-introduction speech. Then, the participant was presented with the explanation. The participant then made a request to the smart speaker. The smart speaker answered the request with the appropriate information. The participant then answered questionnaires. Each participant performed the above procedure a total of 16 times under different conditions. Finally, an interview was conducted. Dialog scenarios for 16 functions of the smart speaker were prepared. The correct request criterion was prepared in advance and a reply of "I'm sorry, I didn't understand that fully." was prepared for request failures. The order of the conditions was counterbalanced across participants.

As an absolute measurement scale, a scale for measuring mental workload [2] was prepared to confirm whether the proposed method affected the user's cognitive load. The mental workload, referred to as Raw TLX, was the arithmetic mean value of six subjective sub-scales that were rated within a 100-point scale. In addition, a scale for measuring the impressions of understandability (7-point Likert scale) were prepared. For relative measurements, the participants were asked to rearrange the $2 \times 2 = 4$ conditions of the two factors (number of agents, type of speech) in decreasing order of "easy to understand". To rank the results, 3 points, 2 points, 1 point, 0 point were allocated in order from 1st place to 4th place.

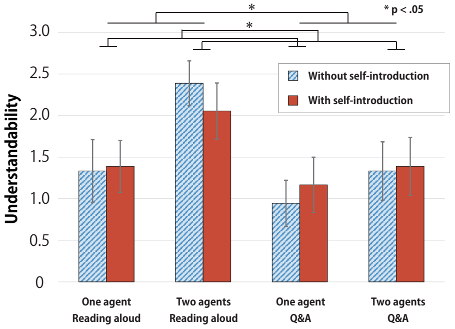

Figure 2: The relative evaluation results of the ease of understanding. The main effects of number of agents and type of speech were significant.

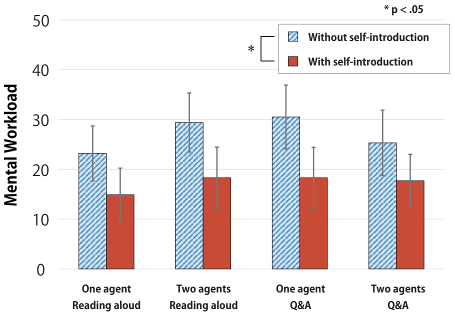

Figure 3: The absolute evaluation results of the mental workload (Raw TLX). The main effect of self-introduction was significant.

Results and Discussion

The relative evaluation results of the ease of understanding are shown in Figure 2. The bar graph represents the averaged values, error bar represents the standard error of mean value. The main effects of number of agents and type of speech were significant as determined by three-way ANOVA, $F(1,64) = 5.631, p < .05, \eta^2 = .07$, $F(1,64) = 5.631, p < .05, \eta^2 = .07$. Figure 2 shows that the style of two agents was easier to understand than the style of one agent. It is presumed that the sub-agent was effective in increasing voice clarity and informality. For example, one participant noted "*I think that it was easy to understand if there were two voices. It was easy to understand because the speech was distinct.* (ID17)" Another participant said "*I felt that the male speech made the instruction less formal.* (ID14)" Figure 2 also shows that the style of reading aloud type was easier to understand than the style of Q&A type. It is presumed that the participants identified with the sub-agent if the sub-agent read aloud specific examples of use. For example, one participant said "*The male utterance is what I would say.* (ID18)" On the other hand, the Q&A format made it harder to understand the instruction. This result differs from the findings of Kubota *et al.* [4]. One reason is that the speech used in the Q&A type was about 10 seconds longer than the speech used for the reading aloud type. Another reason is that the previous research dealt with news information, whereas this paper deals with the limited topic of function explanation. Since the sentence construction was relatively simple, the effectiveness of starting with a presentation of the context by inserting the question was small.

The absolute evaluation results of the mental workload (Raw TLX) are shown in Figure 3. The main effect of self-introduction was significant as indicated by three-way ANOVA, $F(1,64) = 4.87, p < .05, \eta^2 = .07$. Figure 3 shows that the group with self-introduction exhibited lower mental workload than the group without self-introduction. It appears that the self-introduction of paired agents is especially effective in helping the user to understand the explanation. For example, one participant

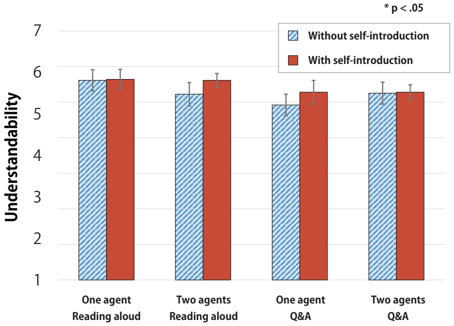

Figure 4: The absolute evaluation results of the ease of understanding. No significant main effects or inter-action terms were found.

of the group with self-introduction said "*I would like the self-introduction if there are two or more people.* (ID18)"

The absolute evaluation results of the ease of understanding are shown in Figure 4. No significant main effects or inter-action terms emerged from the three-way ANOVA results. Figure 4 shows that three factors were not significantly effective. The main effect was significant in the relative evaluation, but not in the absolute evaluation. It is presumed that the factor of interest in the function and the factor of simplicity of the voice command were the major factors in the absolute evaluation. For example, one participant said "*It was hard to understand because the function was new to me.* (ID14)" Another participant said "*It was easy to understand because the voice command was simple.* (ID16)"

CONCLUSIONS

We proposed an instruction presentation method for smart speakers that introduces a vocally-unique sub-agent which performs a different role from the main agent. We verified the effects on understandability of instructions and cognitive load requirements by introducing multiple talking agents that are heard but not seen. Experiments showed that the instructions provided by the main agent and sub-agent were easier to understand than those issued by just the main agent. Experiments also showed that the instructions were most easy to understand when the sub-agent read aloud a specific example. Since our study involved just 18 participants from two groups of nine participants, future work includes conducting the experiment with many more participants.

REFERENCES

[1] A Baylor and Suzanne J Ebbers. 2003. Evidence that multiple agents facilitate greater learning. *Artificial intelligence in education: Shaping the future of learning through intelligent technologies* (2003), 377–379.

[2] Sandra G Hart and Lowell E Staveland. 1988. Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In *Advances in psychology*. Vol. 52. Elsevier, 139–183.

[3] Nobukatsu Hojo, Yusuke Ijima, and Hideyuki Mizuno. 2018. DNN-Based Speech Synthesis Using Speaker Codes. *IEICE TRANSACTIONS on Information and Systems* 101, 2 (2018), 462–472.

[4] H. Kubota, K. Yamashita, T. Fukuhara, and T. Nishida. 2002. POC caster: Broadcasting Agent Using Conversational Representation for Internet Community. *Transactions of the Japanese Society for Artificial Intelligence* 17 (2002), 313–321.

[5] Ewa Luger and Abigail Sellen. 2016. Like having a really bad PA: the gulf between user expectation and experience of conversational agents. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. ACM, 5286–5297.

[6] George A Miller. 1956. The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological review* 63, 2 (1956), 81.

[7] Martin Porcheron, Joel E Fischer, Stuart Reeves, and Sarah Sharples. 2018. Voice Interfaces in Everyday Life. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM, 640:1–640:12.

[8] William Swartout, David Traum, Ron Artstein, Dan Noren, Paul Debevec, Kerry Bronnenkant, Josh Williams, Anton Leuski, Shrikanth Narayanan, Diane Piepol, et al. 2010. Ada and Grace: Toward realistic and engaging virtual museum guides. In *International Conference on Intelligent Virtual Agents*. Springer, 286–300.