A Spoken Dialogue System Using Virtual Conversational Agent with Augmented Reality

Shinji Miyake* and Akinori Ito*
* Graduate School of Engineering, Tohoku University, Sendai, Japan
E-mail: miyake@spcom.ecei.tohoku.ac.jp, aito@spcom.ecei.tohoku.ac.jp Tel: +81-022-717-7085

***Abstract*—We have developed a spoken dialogue system using virtual conversational agent with augmented reality. The proposed system has architecture based on question and answer database that contains many question and answer pairs. Additionally, we have developed two agents displayed using augmented reality, which behave as avatars of objects to be operated. We evaluated user's impression as well as response accuracy of our proposed system. As a result, the existence of an agent increased user's feeling of vividness of conversation and easiness to talk to the system. In addition, the system with an agent showed better response accuracy than the system without agents.**

Index Terms: Spoken dialogue system, Virtual agent, Augmented reality

I. Introduction

Speech is a natural way of interacting with objects. There have been numerous spoken dialogue systems so far [1]. Among them, we are developing a spoken dialogue system that makes it easy for a user to interact with things around us, such as appliances, mobile robots etc [2], [3]. Our target is relatively small-sized dialogue system that can be developed easily. Because such a dialogue has only a few turns to accomplish a task, complicated dialogue control is not necessarily needed. Instead, this kind of system need to be developed rapidly at low cost, and maintenance of the system should be easy.

There have been several systems for the similar purposes such as an agent-based information kiosk [4], [5], [6], smart home control [7] or an interactive communication robot [8]. These systems have a "subject" of dialogue, such as animated character on a display or a body of a robot. Most systems without robots use talking heads or anthropomorphic agents [9], which are virtual partners of the dialogue.

Although most dialogue systems with agents focus on how to build the dialogue system and how to design the behavior of the agent including synchronization of synthesized speech and facial expression, there have been few works that consider how the agent should be displayed. Our goal is to develop a spoken dialogue system for things in usual environments such as TV, air conditioner, microwave oven, etc., which is basically same as the "smart home" application [7]. In this research, we propose a dialogue system as a communication tool between a human and home electric appliances, rather than entering commands correctly. Not only appliances, things in our environment are expected to become more and more intelligent [10], and this virtually almost all object in the environment (including a pencil or a teapot) will have information which we want to draw through spoken dialogue. Our ultimate goal is to realize an environment in which we can interact with almost every object through spoken dialogue.

Now, how should the agent displayed when manipulating a specific object? Conventional systems assume a display and microphone for the dialogue agent anywhere in the room. However, we think that the agent should be displayed at the very position where the target object is. In other words, when manipulating a real object, the user and the agent should share the same space.

In this work, we propose a spoken dialogue system that displays the virtual agent using augmented reality (AR). The AR[11] is a technology that superimposes virtual objects and/or informations into captured video of real world. The AR technology enables to create images that looks as if the objects exist in the real world. By combining the AR technology and the spoken dialogue system with an agent, we can display the agent near the object we want to manipulate. In this paper, we report development of a spoken dialogue system with AR-based virtual agent, and a result of a dialogue experiment.

II. A Dialogue System "P-CE"

A. System Framework

In the proposed dialogue system, we assume that the user makes dialogues using speech using a device such as a smartphone or a tablet PC that are equipped with a camera for capturing the target object. When the user manipulate an object, the user captures that object using the camera. The objects manipulated by dialogue have markers (visual tags) that are used for displaying virtual agent using AR. We prepare different agent for different object, and the agent behaves as an avatar of the object. The basic procedure of the dialogue session is as follows.

1) The user directs the camera of the dialogue device to the object he/she want to manipulate.
2) The captured object and its marker are displayed on the device, and then the agent appears in the display.
3) The user starts dialogue with the agent.
4) After the dialogue, the device communicates with the object using a network and remotely manipulate it.

When making the dialogue, the agent and the object are always displayed together on the device. Therefore, not only verbal conversation, non-verbal expression of the agent is available (such as pointing a switch the user need to push). Figure 1 illustrates the proposed dialogue system. This proposed

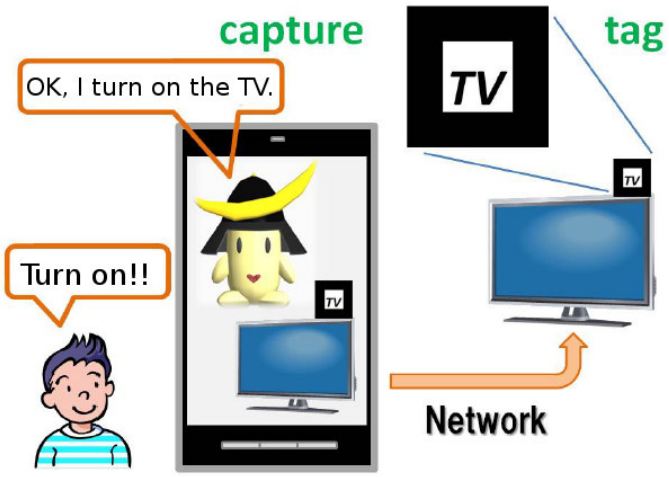

Fig. 1. Overview of the proposed dialogue system

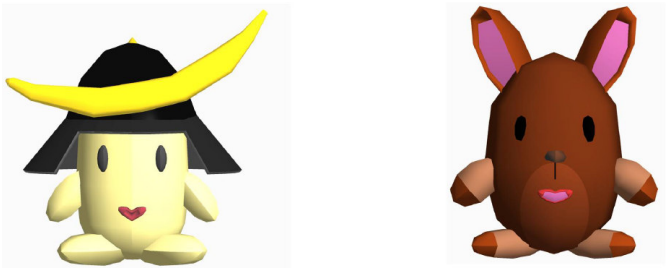

Fig. 2. Prepared agents

approach also can expect to make improvements the acoustic problems of distant talking to appliances.

B. Implementation of the system

We exploited ARToolkit [12] for realization of AR. We prepared two agents for using in the later experiment. Figure 2 shows the prepared agents. Several postures were prepared for one agent to animate the agent. When displaying an agent, we switched two motions (motion for waiting and motion for talking) according to the state of the dialogue. We did not any synchronization between the motion and the synthesized speech. We prepared two markers, and switched the agent according to the recognized marker.

The spoken dialogue system is based on Question-and-Answer database [5]. This system has a database that includes pairs of an assumed question (an example sentence) and an answer to the question. Table I shows examples of the example sentences, and Table II shows the answer sentences. Both an example sentence and an answer sentence are associated with a tag. When an input utterance is given, the utterance is transcribed using a large vocabulary continuous speech recognizer, and converted into a word sequence. Then the word sequence is compared with the example sentences in the database, and the example sentence most similar to the input utterance is chosen. Then the tag of the chosen example sentence is extracted, and the answer sentence that is associated

TABLE I
EXAMPLES OF EXAMPLE SENTENCES

Tag	Sentence
#1001	Turn on the TV
#1001	Power on
#1002	Turn off the TV

TABLE II
EXAMPLES OF ANSWERS

Tag	Sentence
#1001	OK. I turn on the TV.
#1002	OK. I turn off the TV.

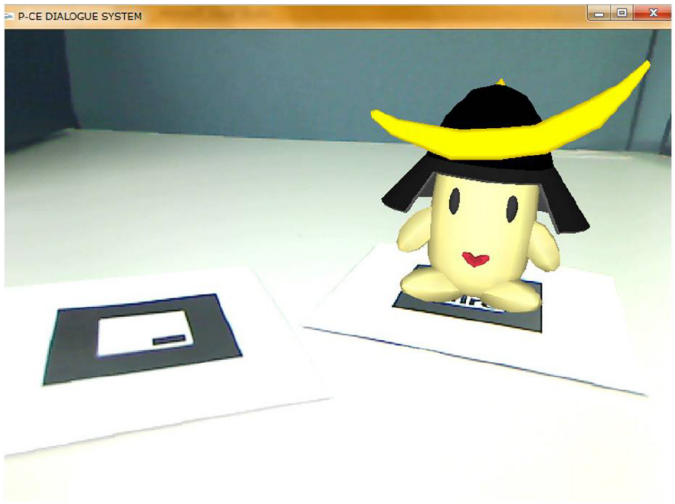

Fig. 3. Screenshot of the dialogue system "P-CE"

with the chosen tag is synthesized. Note that the database is prepared task-by-task, where one task corresponds to a specific object such as a TV or an air conditioner. On implementing the system, we used Julius [13] as a speech recognition engine and AquesTalk2 [14] as a speech synthesizer.

Figure 3 shows the screenshot of the developed dialogue system "P-CE". When the camera captures a marker, the agent associated with that marker is activated and superimposed into the captured image. At the same time, the database for manipulating the captured object is loaded into the spoken dialogue system, and the system starts a dialogue.

III. EVALUATION EXPERIMENT

A. Experimental Conditions

We carried out an experiment to evaluate the proposed dialogue system. We did not use a mobile device but a PC with an LC display, because of the easiness of evaluation.

We prepared two tasks: "TV" and "Air conditioner." We employed 9 participants (8 males and 1 female) who were not familiar with spoken dialogue system. Before the dialogue, we gave a participant an instruction, where five examples of available user utterances were presented, and we instructed the participant to make conversation freely. Number of dialogues per participant was 4 in average.

We conducted two sets of dialogues for one participant, where no agent was displayed in the first set, and the agent was displayed in the second set. Table III shows the number of dialogues. All dialogues were conducted in Japanese. After the

TABLE III
NUMBER OF GATHERED DIALOGUES

	TV	A.C.	Total
w/o agent	48	43	91
with agent	47	45	92

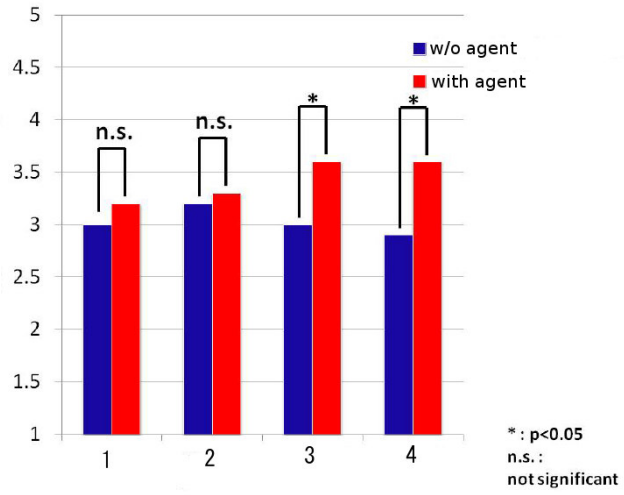

Fig. 4. Average of evaluation score for all conditions

dialogues, we asked the participants to answer a questionnaire that have four evaluation items with five-scale grades (1 to 5, 5 to be the best), as follows.

1) Easiness to use the system in total
2) Smoothness of the conversation
3) Vividness of the conversation
4) Easiness to talk to the system

In addition to the above evaluations, we asked the participants to describe opinions to the system.

B. Result of subjective evaluation

Figure 4 shows average scores of all conditions. The x-axis of the figure corresponds to the four evaluation items of the questionnaire shown above. We conducted statistical test to check whether the differences between the conditions (with or without an agent) were statistically significant. As a result, we obtained significant differences for two questions ("vividness of the conversation" and "easiness to talk to the system"). This result suggests that the agent improves the feeling of "liveliness" or "friendliness" of the conversation.

C. Analysis of user utterances

We analyzed the gathered transcriptions of the dialogues, and noticed that the utterances toward the system with an agent seem to be shorter than those toward the system without an agent. Here are such examples:

Without agent:	*Please turn on the TV switch*
With agent:	*Turn on*

To investigate this phenomenon, we counted length of utterances under each of the conditions. Here we calculated three

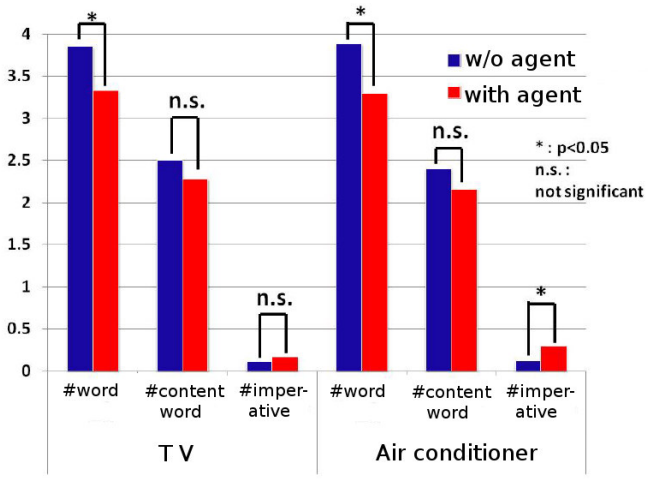

Fig. 5. Analysis result of user utterances

indices: number of words in an utterance, number of content words in an utterance and number of imperative verbs in an utterance. The result is shown in Figure 5. To confirm the difference of the results, we conducted statistical test (t-test). As Figure 5 shows, we found that number of words were significantly different in both TV and air conditioner tasks. Number of content words were not significantly different, which suggests that the difference was mainly number of function words. Function words of Japanese express speaker's attitude, politeness or social relationship in addition to grammatical role, which seems to be the reason of this difference.

The participants described in the questionnaire that they could talk to the agent when the agent was displayed, but when the agent was absent, they did not know how to talk to the system. In this case, the utterances became "conservative" (politer, longer, and less omission), and more repetitions were also observed. This was a reason of longer utterances.

D. Comparison of response accuracy

We compared the response accuracy under each condition. response accuracy means accuracy of response of the system for the user utterances. As the utterance is proved to change according to existence of an agent, response accuracy also might be affected by an agent. On matching the transcription of the input utterance, we examined three methods of matching related to Japanese grammar. A Japanese verb inflects according to mood of the sentence. Here are examples:

(A) *Onryou o agete kudasai*	Increase volume (polite)
(B) *Onryou agero*	Increase volume (impolite)

Here, *agete kudasai* and *agero* have the same meaning "increase" but they are in different forms (*agete* and *agero*) because of difference of mood. As we match the input utterance to the database using surface form of the words, *agete* and *agero* in the above example will be regarded as different words. To match verbs in different mood, we compared the following three methods of matching.

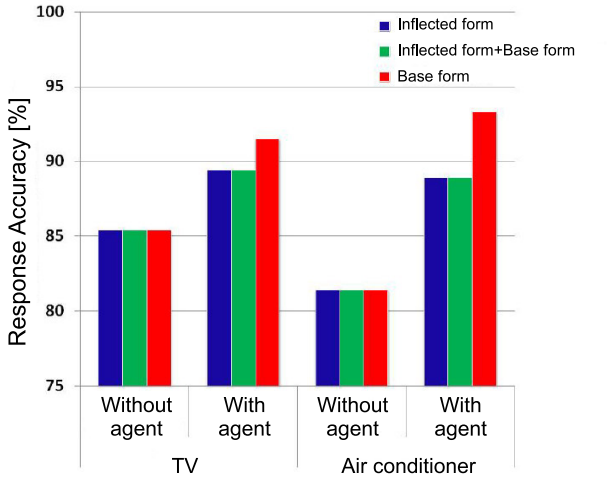

Fig. 6. Response accuracies for all conditions

- Use the surface form of the words (the conventional method).

  (A) *onryou o agete kudasai*
  (B) *onryou agero*

- Use the combination of the surface form and the base form.

  (A) *onryou+onryou o+o agete+ageru kudasai+kudasaru*
  (B) *onryou+onryou agero+ageru*

  Using this method, we can distinct different words that have the same surface form.

- Use the base form.

  (A) *onryou o ageru kudasaru*
  (B) *onryou ageru*

  Using this method, we can match the same words with different surface form.

The experimental result is shown in Figure 6. This result shows that the existence of an agent improves the response accuracy. The reason of this improvement is not completely understood, but main reason seems to be less repetition, less disfluency, and as a result, less recognition rate degradation. Therefore, we think that existence of the agent might have affected the user utterances.

Result from all subjects first tried the system without the agent, the participants might have learned how to use the system. But as expected, matching using base form improved the accuracy only when an agent was displayed.

IV. Conclusions

In this paper, we proposed a spoken dialogue system to interact with things in an environment using agents with AR. The main idea of this system is to attach markers to the objects to be manipulated with the system. When a user want to manipulate an object, the user captures the object with a camera, and then the marker is recognized by the AR system and the agent of that object appears on the screen of the device. The user makes conversation with that agent to manipulate the object or obtain information from the object. The experimental result showed an interesting fact that the existence of an agent affected the user utterance and the effect was positive from spoken dialogue system point of view (vividness of the conversation for the objects, not a command, etc).

In the future work, we need to evaluate the effect of different agent design on the user evaluation and system performance [15], [16]. In addition, we need to develop a system with which a developer can develop more realistic dialogue more easily.

V. Acknowledgment

Part of this work was supported by Grant-in-Aid for challenging Exploratory Research by Japan Society for the Promotion of Science (JSPS), No. 24652111.

References

[1] M. F. McTear, "Spoken dialogue technology: enabling the conversational user interface," *ACM Computing Surveys*, vol. 34, no. 1, pp. 90–169, 2002.
[2] T. Konashi, M. Suzuki, A. Ito and S. Makino, "A spoken dialog system based on automatic grammar generation and template-based weighting for autonomous mobile robots," In *Proc. Int. Conf. on Spoken Language Processing*, (2004), CD-ROM.
[3] S. Hahm, A. Ito, K. Awano, M. Ito and S. Makino, "Utterance Classification for Combination of Multiple Simple Dialog Systems," In *Proc. IEEE Int. Symp. on Parallel and Distributed Processing with Applications Workshop*, pp. 171–176, 2011.
[4] J. Gustafson, N. Lindberg and M. Lundeberg, "The August spoken dialogue system," in *Proc. Eurospeech*, 1999.
[5] R.Nisimura, A. Lee, H. Saruwatari and K. Shikano, "Public Speech-Oriented Guidance System with Adult and Child Discrimination Capability," in *Proc. ICASSP*, vol. I, pp. 433–436, 2004.
[6] S. Kopp, L. Gesellensetter, N. C. Krämer and I. Wachsmuth, "A Conversational Agent as Museum Guide – Design and Evaluation of a Real-World Application," LNCS, vol. 3661/2005, pp. 329-343, 2005.
[7] C. Kühnel, B. Weiss and S. Möller, "Talking heads for interacting with spoken dialog smart-home systems," in *Proc. Interspeech*, pp. 304–307, 2009.
[8] J. Osada, S. Ohnaka and M. Sato, "The Scenario and Design Process of Childcare Robot, PaPeRo," in *Proc. Int. Conf. on Advances in Computer Entertainment Technology,* pp. 80–87, 2006.
[9] S. Kawamoto et al., "Open-source Software for Developing Anthropomorphic Spoken Dialog Agents," in *Proc. PRICAI-02, Int. Workshop on Lifelike Animated Agents,* 2002.
[10] L. Atzoria, A. Ierab and G. Morabito, "The Internet of Things: A survey," *Computer Networks*, vol. 54, no. 15, pp. 2787–2805, 2010.
[11] R. T. Azuma: "A Survey of Augmented Reality," *Presence: Teleoperators and Virtual Environments*, vol. 6, no. 4, pp. 355–385, 1997.
[12] H. Kato, M. Billinghurst, S. Weghorst and T. Furness, "A Mixed Reality 3D Conferencing Application," *Human Interface Technology Laboratory*, University of Washington, 1999.
[13] A. Lee, T. Kawahara and K. Shikano, "Julius — An Open Source Real-Time Large Vocabulary Recognition Engine," in *Proc. Eurospeech*, pp. 1691–1694, 2001.
[14] A-Quest, http://www.a-quest.com/
[15] H. McBreen and M. Jack, "Empirical Evaluation of Animated Agents in a Multi-Modal E-Retail Application," In *Proc. AAAI Fall Symp. on Socially Intelligent Agents,* 2000.
[16] N. C. Krämer, N. Simons and S. Kopp, "The Effects of an Embodied Conversational Agent's Nonverbal Behavior on User's Evaluation and Behavioral Mimicry," *Intelligent Virtual Agents, LNCS*, vol. 4722/2007, pp. 238–251, 2007.