Semi-autonomous Guide Agents with Simultaneous Handling of Multiple Users

Yusuke Muraki, Haruki Kawai, Kenta Yamamoto, Koji Inoue, Divesh Lala and Tatsuya Kawahara

Abstract In this paper we describe a laboratory guide dialogue system where multiple users interact with their own semi-autonomous agent. When the agent cannot answer a user's question, it is able to hand control of the dialogue to a single remote human operator who is monitoring the conversations. We describe how this system functions even in cases where different agents need to hand over control to the operator simultaneously. We also conduct a subjective experiment which showed that our multiple-user system is not significantly different than a single-user system while also performing better than a fully autonomous system.

1 Introduction

Spoken dialogue systems have recently been improved, particularly due to advances in natural language processing and automatic speech recognition (ASR). However, since errors in these two processes still exist and other phenomena such as turn-taking have to be carefully considered, completely autonomous systems such as

Yusuke Muraki
Kyoto Univerity Graduate school of Informatics, e-mail: muraki.yusuke.ac@gmail.com

Haruki Kawai
Kyoto Univerity Graduate school of Informatics, e-mail: kawai@sap.ist.i.kyoto-u.ac.jp

Kenta Yamamoto
Kyoto Univerity Graduate school of Informatics, e-mail: yamamoto@sap.ist.i.kyoto-u.ac.jp

Koji Inoue
Kyoto Univerity Graduate school of Informatics, e-mail: inoue@sap.ist.i.kyoto-u.ac.jp

Divesh Lala
Kyoto Univerity Graduate school of Informatics, e-mail: lala@sap.ist.i.kyoto-u.ac.jp

Tatsuya Kawahara
Kyoto Univerity Graduate school of Informatics, e-mail: kawahara@i.kyoto-u.ac.jp

conversational agents are still not at the level of human understanding, particularly in open-domain systems.

Even for question-answering (Q-A) spoken dialogue systems, it is unknown if they will ever be able to completely replace humans in all domains. Some tasks may require large language models or a handcrafted approach, but even with these it is still possible that the system fails at correctly handling, out-of-domain requests. To provide support for these situations, our approach in this work is to make the system semi-autonomous to handle simple requests, while offloading unmanageable conversation to a human operator.

Such an approach may seem redundant - if a human can intervene then it does away with the need for a dialogue system in the first place. However, we consider the problem of mass usage of a spoken dialogue system which has to address the needs of *multiple users* interacting with their own agents. We propose as an example scenario a museum agent where many users wish to ask the agents details about an exhibit. If an agent encounters a question which is not answerable, one solution could be for a remote operator who is listening to the conversation to quickly intervene. However, if there are many such agents, the number of operators would need to scale with the number of users. To reduce resources we propose using just one human operator to monitor multiple dialogues, with the ability to efficiently intervene and deal with any requests that the system cannot handle.

This presents a range of problems, a major one being how the human can know when to intervene. If the operator is listening to only one conversation this decision may not be difficult, but it poses problems when dealing with multiple people, particularly since an operator cannot listen to multiple conversations simultaneously. Our solution is for the system itself to know when the operator should take over and notify them so that they can efficiently handle the user's request.

In this work, we propose such a system in the context of a virtual laboratory guide who answers user questions. The system will handle questions that it can answer, but when it cannot it notifies a human operator who will speak directly with the user. The operator can manage three different users in such a manner, including dealing with multiple simultaneous requests. Figure 1 shows the general concept of the system.

The goal of this system is to show that this simultaneous parallel architecture can adequately function in the context of question-answering. In this work we compare our multi-user approach to a fully autonomous system and a single-user system. The dialogue system and user interface uses the Japanese language.

2 Related Work

Tele-operated systems for robots have been previously implemented to aid in elderly care and child socialization [18, 9, 3] and recent field work in a public setting which attempts to disguise the human operator as a robot in a Wizard-of-Oz manner [1, 17].

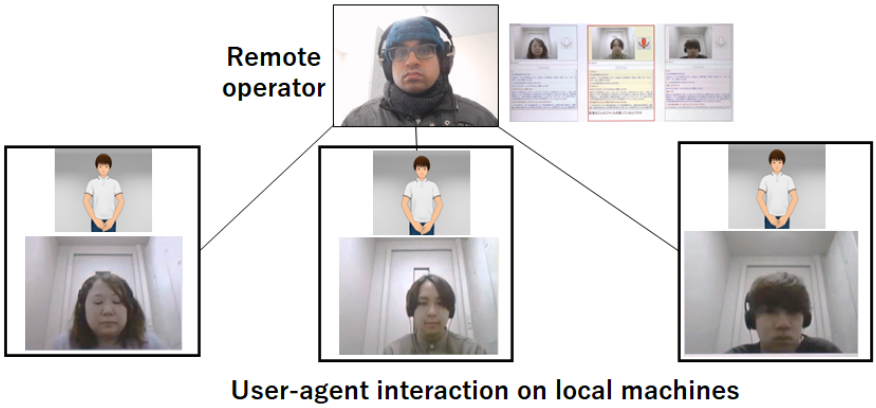

Fig. 1 General concept of the semi-autonomous multi-user system.

Tele-operation of multiple agents has also been addressed, but with most focus on mobile robots rather than conversation [11, 15, 8].

In this work our goal is to use a remote operator to address situations where the agent cannot handle out-of-domain questions. Previous literature has dealt with this issue by improving the dialogue system itself [16, 12, 7]. The approach of handing over control to a human operator has been proposed in other work, however this has been mainly for text-based chatbots rather than spoken dialogue [13, 14] so is not directly applicable to our case which contains additional issues such as incorrect speech recognition, fillers and lack of punctuation. Furthermore, the operator interface described in [14] relies heavily on the operator being able to efficiently process a large amount of textual information and generate an appropriate response. In our system, the relevant question is displayed and highlighted so that much of the cognitive load is towards reading the question and generating an answer. Recovery handover from agents to humans in spoken dialogue is rare even for dyadic interaction [2].

Perhaps the closest work to ours is by [5], who implemented a system in which multiple social robots were controlled simultaneously. However operator interventions involved selecting from a set of utterances, rather than being able to directly converse with the user. Furthermore, the operator must constantly monitor the robots for communication breakdowns. In our system the agent is able to identify when to hand over control to the operator, who directly speaks to the user.

3 Laboratory Guide Dialogue System

The agent used in this task is a version of MMDAgent-Ex, a toolkit for voice interaction with a humanoid avatar [10]. We use a text-to-speech system to generate

utterances. These utterances are sent to the agent which then provides lip synchronization. Automatic speech recognition (ASR) is implemented by a sub-word unit recurrent neural network with an attention mechanism.

The dialogue system itself is quite simplified. The agent begins with a short self-introduction and then afterwards explains the research activities that occur in the laboratory. After each system turn, the user is given an opportunity to ask any question, even if not related to the agent's talk. In this work, we adopt a simple Q-A database system, but it is not possible to cover all possible questions and so we assume that at some time the user will ask an out-of-domain topic which cannot be handled by the automated dialogue system. If the user does not ask anything and is silent, the agent continues the dialogue.

Firstly, we implement a question detection module by using an existing two-layer LSTM neural network that was trained with annotated data on a human-robot interaction corpus [6] with an F1 score of 84.4%. The model is a binary classification of whether the input sentence is a question or not. If it is a question, the system tries to find the corresponding answer from the database using Elasticsearch [4]. If the corresponding answer can be found, it is given to the user otherwise the question is further classified into one of two types: *Subjective* or *Objective*. We define subjective as questions that are opinion-based, such as "Why do you like the lab?", whereas objective questions are those which a person would be able to answer quickly, such as "What is your age?".

We use this classification of the question to decide how to prioritize which questions the operator needs to take control of the system and answer as quickly as possible. Our approach is to base this prioritization on the question's subjectivity. Our rationale is that subjective questions should be prioritized over objective ones because these require an opinionated response and so naturally it is expected the agent can answer this with minimum thought. On the other hand, objective questions may require more consideration by the agent and in some cases it is acceptable if the agent defers to a later time.

To train this model we collected samples by asking crowd-sourced workers to create subjective and objective questions for five situations, including laboratory and travel guide scenarios. We then trained our model based on these questions. In total, the workers provided 1,806 question sentences, and we trained an LSTM model on these using 25% of the questions as test data. We obtained F1 scores of 76.1% and 79.5% for the subjective and objective questions, respectively. This model will be used to decide how to handle operator switching, as described in Section 4.2.

4 Operator System

The operator system is designed to allow the operator to monitor multiple users at the same time and to quickly understand the state of the conversation, including preparation for taking control from the system. Here we describe the architecture in the context of the laboratory guide dialogue system, however it should be noted

that many of the design is applicable for any dialogue system, requiring only a few modifications. Therefore we propose that this system is reasonably agnostic to the dialogue system being employed and even the type of agent used.

We also compared our interface to that of a handover system using chatbots [14], where the operator receives a large amount of information at handover time, and has to consider this to provide an appropriate answer. The reason for this is the use case of a customer service agent, where it is important to know for example the customer's history, purchase date of items and category of problem. In our scenario, we are more interested in quickly answering questions, so the most relevant information for the operator is a notification of when the handover is about to occur and the relevant question that needs to be answered. Therefore we designed the interface with these two requirements in mind.

The operator's GUI is presented as an image to the right of the operator in Figure 1. It contains video displays showing each user and their conversation history, updated when an ASR result or agent response is received. By clicking on the panel display, the operator can directly listen to the conversation between a user and agent. Clicking on the microphone icon allows the operator to toggle if they take over from the system and speak directly to the user. A question display is used for automatic switching. We show an example of a panel for an individual user in Figure 2.

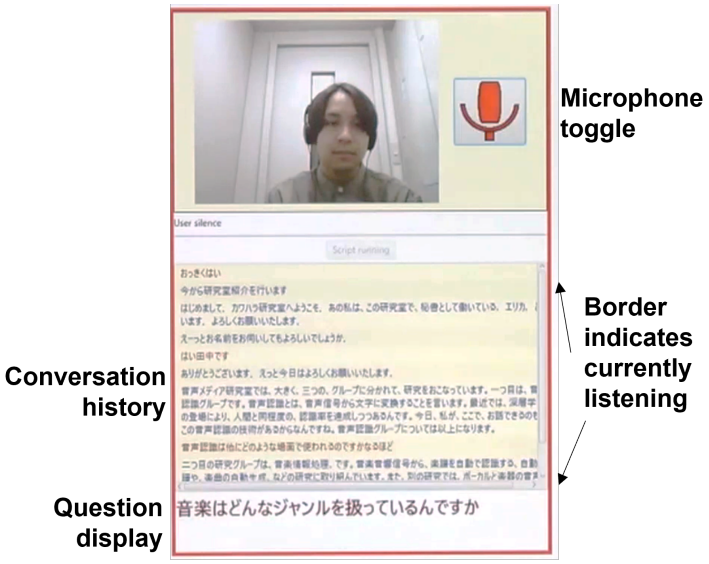

Fig. 2 Operator system GUI for an individual user

4.1 Automatic switching

The dialogue system informs the operator when it cannot answer a question that a user has asked. If an unanswerable question has been received, the operator system begins automatic switching. This means the operator will be required to take control. A message is displayed to the operator notifying that a takeover will begin, along with the question that the user has asked. The operator then has 4 seconds to prepare themselves for the takeover.

The system will eventually hand control of the conversation to the operator, and anything spoken into the operator's microphone will be transmitted to the user via the agent, with its lip synchronization matching the operator's voice. Once the operator has finished their talk, they toggle the microphone off to switch control back to the system, which continues with the next dialogue.

4.2 Simultaneous request handling

Our system extends the use case of one user to multiple users to handle simultaneous requests. If an operator is "busy", then the system must decide how to deal with the user in a manner that lets them continue the conversation smoothly.

First, the system classifies the user's question as subjective or objective. If it is subjective, then the operator should deal with this as soon as they can, so the system uses a filler (e.g. "Let me see...") to buy time for the response. An example interaction is below:

User What kind of lab activities do lab members like to do?
System Let me see...
Operator-controlled I think they enjoy going out for dinner together and playing sports.

The goal is to keep the delay between the system and the operator-controlled responses as short as possible, preferably under 5 seconds so that the user does not feel impatient. If multiple users' questions are queued up, the switches between them are done in a FIFO order to try and minimize this delay.

If the question is objective, the system will inform the user that they will answer the question at a later time. Our rationale for this is that objective information can be delayed as it does not seriously hinder the flow of the conversation. The system continues its explanation, with the unanswered question displayed in the operator GUI. During the dialogue, the operator can *manually* take control by clicking on the microphone icon and answering the question. Multiple questions can be stored in this manner and answered at a later time. An example of this interaction is as follows:

User What are the usual lab hours?

System Oh, I'll answer that question soon. Let me continue with the explanation first.
guide explanation continues as normal
Operator-controlled Going back to that question you asked, lab hours are usually 9 to 5, but you are mostly free to work whenever you want.

The operator decides when to address this question at a later time. In our scenario, there is a time for open questions at the end of the guide's talk, so this is a convenient period in which these questions can be addressed without the feeling that the operator has "barged in" to the conversation.

The overall process for handling questions is shown in Figure 3.

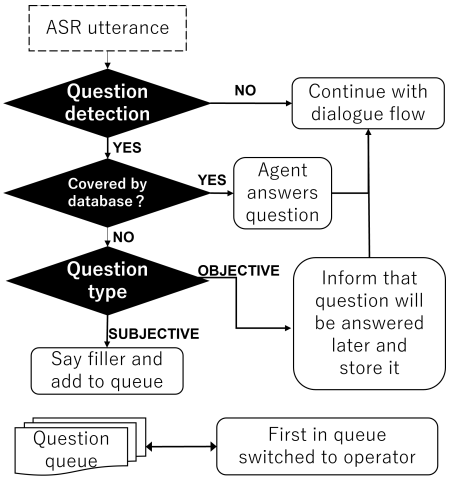

Fig. 3 Flow diagram for question request handling

4.3 System simulation

We also wanted to confirm that our approach is efficient with multiple simultaneous users. We conducted a simulation to estimate the average waiting time per user (disregarding operator switching time) over a session with varying numbers of users. Four utterances per user were simulated across the session. We make an assumption that 75% of a user's utterances are questions and 50% of these are subjective. We then adjusted the coverage of the Q-A database and compared this to a baseline system where users wait for their request to be answered, similar to a call center.

Simulation results are shown in Table 1. We predict our system will reduce waiting time compared to the baseline. If the database has high coverage (75%) the difference between the baseline and proposed system is minimal, since the agent

can answer most questions. On the other hand, if coverage is low (25%) with more need for human intervention, then the proposed system greatly reduces the waiting time compared to the baseline.

	Baseline (# users)			Proposed (# users)		
	3	**5**	**7**	**3**	**5**	**7**
25%	15.3	21.4	31.5	6.8	9.3	13.5
50%	6.2	12.9	16.7	4.5	2.8	4.1
75%	2.1	2.2	3.1	1.3	2.1	2.2

Table 1 Estimated waiting time per person based on system simulation. Left column indicates Q-A database coverage.

5 Experiment

The goal for this experiment is to subjectively compare our proposed system against two other systems. The first is a fully autonomous system where if the system cannot answer a question it explicitly states that it is unable to do so, and continues the dialogue. The second system is a dyadic version of the proposed system with the operator only having to handle one subject. We assume it will be cognitively easier than our proposed system, which deals with multiple subjects. However, our intention is for the proposed system to be comparable to justify our multi-user approach.

All subjects used the same laboratory guide dialogue system. They were informed beforehand that they could ask questions at any time. Subjects who had a human operator were also fully informed beforehand that an operator was monitoring the conversation and could intervene during the dialogue. For our proposed system, two to three subjects participated with the system at the same time. We had 10 sessions of these and 28 subjects in total. In addition, 34 subjects used the fully autonomous system while 13 subjects used the dyadic system. The human operator was one of the authors of this work and was the same for all conditions. After the interaction the subjects answered the below questions on a five point Likert scale.

- I understood the agent's explanation.
- The dialogue was natural.
- The agent's responses were appropriate.
- The interaction proceeded smoothly.
- The agent's actions were appropriate during question handling.
- I received the answer I was looking for.

6 Results

We conducted t-tests ($\alpha = 0.05$) to compare the proposed system against the two other conditions. Results are shown in Figure 4. We find no significant difference between the proposed and dyadic system for any metric, although the number of subjects who undertook the dyadic condition was significantly less. The proposed system is favored by subjects in the metrics of appropriateness, smoothness and receiving a satisfactory answer compared to the fully autonomous system. The results suggest that the answers generated from the fully autonomous system are less satisfactory than the proposed system, likely due to the fact that it cannot cover all of the user's questions.

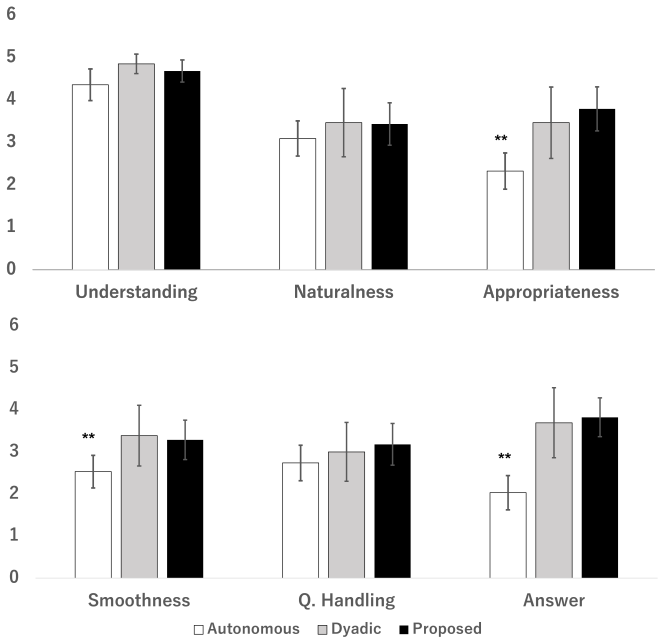

Fig. 4 Mean and confidence intervals for evaluation metrics. ** indicates $p \leq 0.01$ with reference to proposed system.

7 Discussion

Our results show that the proposed system is comparable to a dyadic system and outperforms the fully autonomous system on some metrics. The first result shows that the handling of multiple users does not negatively affect the system, suggesting

that the added delay in answering questions from multiple users at once was either not observed or not perceived as a problem.

Although the proposed system outperformed the fully autonomous system, it was not significantly better at handling questions, perhaps suggesting that users thought the actual method to handle questions was not different, but the answers provided were naturally better coming from the operator. This result also does not consider the performance of the underlying dialogue system itself. From our simulation, we assume a poor dialogue system would benefit more from being able to hand control to a human operator.

We attempt to classify subjective and objective questions in this work, and use this as the basis for the speed of the answer. However, this correlation is not always desirable. For example, an objective question such as "Do you drink coffee" should be answered quickly, but the objective question "How many times last month did you drink coffee?" probably does not need to be answered immediately. A more accurate model would classify based on the urgency of the request, and this needs extra annotation.

Although this paper focuses on the user experience, we also intend to assess how an operator can successfully manage this system in terms of cognitive load. We propose that the interface is basic enough that it does not require much training and the operator has no need to constantly monitor multiple interactions. We also intend to expand the range of conversational scenarios that can be used for the system and evaluate its performance.

Additionally, this paper focuses on automatic switching for one particular scenario, answering out-of-domain questions. We intend to extend the concept of automatic switching to a remote operator to situations such as the user being disengaged or system responses becoming too repetitive. This expands the scope of our system to a wide range of dialogue scenarios.

There are still several issues for improvement of our system. The major one is that currently the voice of the agent is different than that of the operator, which may be somewhat obtrusive for the user. In the experiment the users were informed that there was a human operator who would intervene so that they would not be surprised. We have also not yet analyzed the system in terms of the operator, in particular the cognitive load needed for dealing with multiple conversations simultaneously. Similarly, it is unknown how much this system can scale. We have conducted experiments with 3 simultaneous users but only simulations with up to 7 users. These improvements will be addressed in future work.

8 Conclusion

In this work we present a conversation system in which a remote human operator monitors multiple virtual laboratory guide agents simultaneously. The agents can identify when assistance is required to answer questions from the user, and then hand over control to the operator. We conducted an experiment showing that our

multiple agent system is comparable to a single-agent system and outperforms a fully autonomous system which cannot satisfactorily cover all user questions.

Acknowledgements This work was supported by JST, Moonshot R&D Grant Number JPMJPS2011.

References

1. Baba, J., Sichao, S., Nakanishi, J., Kuramoto, I., Ogawa, K., Yoshikawa, Y., Ishiguro, H.: Teleoperated robot acting autonomous for better customer satisfaction. In: Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems, pp. 1–8 (2020)
2. Benner, D., Elshan, E., Schöbel, S., Janson, A.: What do you mean? A review on recovery strategies to overcome conversational breakdowns of conversational agents. In: International Conference on Information Systems (ICIS) (2021)
3. Chen, L., Sumioka, H., Ke, L., Shiomi, M., Chen, L.: Effects of teleoperated humanoid robot application in older adults with neurocognitive disorders in taiwan: a report of three cases. Aging Medicine and Healthcare **11**, 67–71 (2020)
4. Elastic: Elasticsearch. https://github.com/elastic/elasticsearch/releases (2022)
5. Glas, D.F., Kanda, T., Ishiguro, H., Hagita, N.: Teleoperation of multiple social robots. IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans **42**(3), 530–544 (2011)
6. Kawahara, T.: Spoken dialogue system for a human-like conversational robot ERICA. In: IWSDS (2018)
7. Khan, O.Z., Sarikaya, R.: Making personal digital assistants aware of what they do not know. In: INTERSPEECH, pp. 1161–1165 (2016)
8. Khasawneh, A., Rogers, H., Bertrand, J., Madathil, K.C., Gramopadhye, A.: Human adaptation to latency in teleoperated multi-robot human-agent search and rescue teams. Automation in Construction **99**, 265–277 (2019)
9. Kuwamura, K., Nishio, S., Sato, S.: Can we talk through a robot as if face-to-face? long-term fieldwork using teleoperated robot for seniors with alzheimer's disease. Frontiers in psychology **7**, 1066 (2016)
10. Lee, A., Oura, K., Tokuda, K.: Mmdagent - a fully open-source toolkit for voice interaction systems. In: 2013 IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 8382–8385. IEEE (2013)
11. Lee, D., Franchi, A., Son, H.I., Ha, C., Bülthoff, H.H., Giordano, P.R.: Semiautonomous haptic teleoperation control architecture of multiple unmanned aerial vehicles. IEEE/ASME transactions on mechatronics **18**(4), 1334–1345 (2013)
12. Liang, K., Chau, A., Li, Y., Lu, X., Yu, D., Zhou, M., Jain, I., Davidson, S., Arnold, J., Nguyen, M., et al.: Gunrock 2.0: A user adaptive social conversational system. arXiv preprint arXiv:2011.08906 (2020)
13. Poser, M., Hackbarth, T., Bittner, E.A.: Don't throw it over the fence! toward effective handover from conversational agents to service employees. In: International Conference on Human-Computer Interaction, pp. 531–545. Springer (2022)
14. Poser, M., Singh, S., Bittner, E.: Hybrid service recovery: Design for seamless inquiry handovers between conversational agents and human service agents. In: Proceedings of the 54th Hawaii International Conference on System Sciences, p. 1181 (2021)
15. Shahbazi, M., Atashzar, S.F., Patel, R.V.: A systematic review of multilateral teleoperation systems. IEEE transactions on haptics **11**(3), 338–356 (2018)
16. Shrivastava, A., Dhole, K., Bhatt, A., Raghunath, S.: Saying no is an art: Contextualized fallback responses for unanswerable dialogue queries. arXiv preprint arXiv:2012.01873 (2020)

17. Song, S., Baba, J., Nakanishi, J., Yoshikawa, Y., Ishiguro, H.: Teleoperated robot sells toothbrush in a shopping mall: A field study. In: Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems, pp. 1–6 (2021)
18. Yamazaki, R., Nishio, S., Ogawa, K., Matsumura, K., Minato, T., Ishiguro, H., Fujinami, T., Nishikawa, M.: Promoting socialization of schoolchildren using a teleoperated android: an interaction study. International Journal of Humanoid Robotics **10**(01), 1350,007 (2013)