Advanced Robotics

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/tadr20

Semi-autonomous avatar enabling unconstrained parallel conversations –seamless hybrid of WOZ and autonomous dialogue systems–

Tatsuya Kawahara, Naoyuki Muramatsu, Kenta Yamamoto, Divesh Lala & Koji Inoue

To cite this article: Tatsuya Kawahara, Naoyuki Muramatsu, Kenta Yamamoto, Divesh Lala & Koji Inoue (2021) Semi-autonomous avatar enabling unconstrained parallel conversations –seamless hybrid of WOZ and autonomous dialogue systems–, Advanced Robotics, 35:11, 657-663, DOI: 10.1080/01691864.2021.1928549

To link to this article: https://doi.org/10.1080/01691864.2021.1928549

Published online: 19 May 2021.

ADVANCED ROBOTICS
2021, VOL. 35, NO. 11, 657–663
https://doi.org/10.1080/01691864.2021.1928549

SHORT PAPER

OPEN ACCESS Check for updates

Semi-autonomous avatar enabling unconstrained parallel conversations –seamless hybrid of WOZ and autonomous dialogue systems–

Tatsuya Kawahara[a], Naoyuki Muramatsu[b], Kenta Yamamoto[a], Divesh Lala[a] and Koji Inoue[a]

[a]Graduate School of Informatics, Kyoto University, Kyoto, Japan; [b]Undergraduate School of Informatics and Mathematical Science, Kyoto University, Kyoto, Japan

ABSTRACT

Many people are now engaged in remote conversations for a wide variety of scenes such as interviewing, counseling, and consulting, but there is a limited number of skilled experts. We propose a novel framework of parallel conversations with semi-autonomous avatars, where one operator collaborates with several remote robots or agents simultaneously. The autonomous dialogue system mostly manages the conversation, but switches to the human operator when necessary. This framework circumvents the requirement for autonomous systems to be completely perfect. Instead, we need to detect dialogue breakdown or disengagement. We present a prototype of this framework for attentive listening.

ARTICLE HISTORY
Received 27 February 2021
Revised 13 April 2021
Accepted 27 April 2021

KEYWORDS
Semi-autonomous dialogue; parallel conversations; conversational avatar; spoken dialogue system; attentive listening

1. Introduction

Conversation agents are now prevailing in smartphone assistants and smart speakers, providing many services as well as chatting functions. However, they are regarded as machines or virtual agents at most. Their conversation style is much different from real human communication. This is true for a large majority of communicative social robots. Their dialogue behaviors are also quite different from those of human interactions. An example of this is that we do not speak so long with them as they do not respond in real time. In this context, we have been developing an intelligent conversational android ERICA to be engaged in human-level dialogue [1]. At the moment, it can perform attentive listening with senior people for five to seven minutes, but subjective evaluations suggest that the quality of dialogue is still behind that of a human (Wizard of Oz) dialogue [2]. There are many remaining challenges ahead before the realization of truly human-level conversational robots.

Meanwhile, due to COVID-19, many of us are forced to communicate remotely using a video conference platform, which is sometimes combined with avatar software. While this setting provides a limited modality in communication, it clears away the physical distance or spatial constraints; now we do not have to travel long distances for meetings and conventions. This advantage will prevail even after COVID-19 recedes and become a new normal. For example, people with some constraints, including disabled people or those who need to take care of children or seniors, can serve customers while staying home; Doctors or counselors can see patients remotely.

However, the time constraints will still remain; one person can serve only one customer or patient at a time. With limited human resources in the future, it would be preferable if one person can serve many people simultaneously. This increase in productivity will also theoretically result in more leisure time. In order to make this happen, namely multiple parallel services, we need to automate some part of them. When the service involves dialogue, we need to incorporate an autonomous dialogue system using an avatar.

This framework provides a new perspective to spoken dialogue system research. The systems do not have to provide perfect performance or human-level experiences. Instead, they can turn to real humans when necessary. It is regarded as a hybrid of WOZ and autonomous dialogue systems, but they need to be switched promptly and seamlessly. For example, these systems will give scripted explanations or respond to typical questions, and switch to humans for handling difficult requests or building personal relationships. With the advancement of the autonomous system and an efficient switching mechanism, the proposed framework allows for handling multiple users at one time. In an extremely simple scenario, we can deal with five customers at a time if we automate 80% of the dialogue service. This is the goal of our project,

CONTACT Koji Inoue inoue@sap.ist.i.kyoto-u.ac.jp

'semi-autonomous avatar enabling unconstrained parallel conversations' under the Moonshot R&D program. Here we want to achieve a human-level experience, which is equivalent to talking to a human, for all customers.

2. Concept of semi-autonomous avatar

2.1. System design

The proposed system architecture is depicted in Figure 1. In this framework, one operator serves many remote users in parallel using an avatar, which can be a robot or virtual agent. As users can be in noisy places such as a shop and a hall, speech needs to be enhanced while detecting the environment. When the user is talking with the operator, his/her speech is directly passed to the operator. Otherwise, the speech is processed with the autonomous system (blue box). First, it is transcribed by the automatic speech recognition module, and then its content is extracted by the natural language understanding module. When an appropriate response is generated, it is output via the speech synthesis module. This flow is essentially the same as the conventional dialogue system, but the major difference is the user is talking to a human, so the speaking style would be more similar to real human communication. There may be many utterances in one turn, so real-time backchannels are generated to keep the user engaged in the dialogue. Another major difference is the system can switch to the operator when it cannot handle the user's request appropriately. If the operator is not available immediately, the system still needs to keep the dialogue flowing by chatting.

2.2. Application tasks

There are many potential applications of the system. One task is a presenter or guide. When we make a presentation of research or products in a convention booth or a poster session, we need to take care of many visitors. This setting is easily converted to a virtual platform that allows for remote online presentation, which becomes a norm under COVID-19. It is not efficient to talk one on one, but it is also not possible to turn everything into an autonomous system, so the hybrid system will be useful. It is applicable to a guide in a virtual museum or a tourist spot allowing remote access. We have previously developed such a lab guide system [3].

The second task is attentive listening or counseling. Under COVID-19, many types of counseling are conducted online as many people with troubles and stresses would be relieved by just being listened to by someone. Attentive listening is also set up for senior people for maintaining communication skills and refreshing memories. We have developed an autonomous attentive listening system using the android ERICA, which can take a majority of the role of this task [2]. Thus, it can be extended to the proposed system allowing multiple users to talk in parallel.

The third task is an interview. Under COVID-19, many job interviews and some college admission interviews are conducted online. While it is not realistic to make them handled by an autonomous system, we have developed an autonomous interview system using the android ERICA, mainly for practice or rehearsal [4]. The proposed system, the hybrid of the autonomous system and a human interviewer, allows for efficient interviews handling many applicants in parallel.

The fourth task is consulting. Under COVID-19, many kinds of consulting are also conducted online, while only simple tasks can be done with AI chatbots. Consulting requires expert domain knowledge such as finance, housing and travel. It is expected that these systems can be semi-automated through the advancement of AI and using the proposed system. There are a variety of dialogue-based customer services, but they are similar to consulting combined with a form of presentation. The characteristics of these tasks are summarized in Table 1. Note that the proposed system is expected to be particularly useful when the text-based communication is inconvenient and real-time interaction is critical.

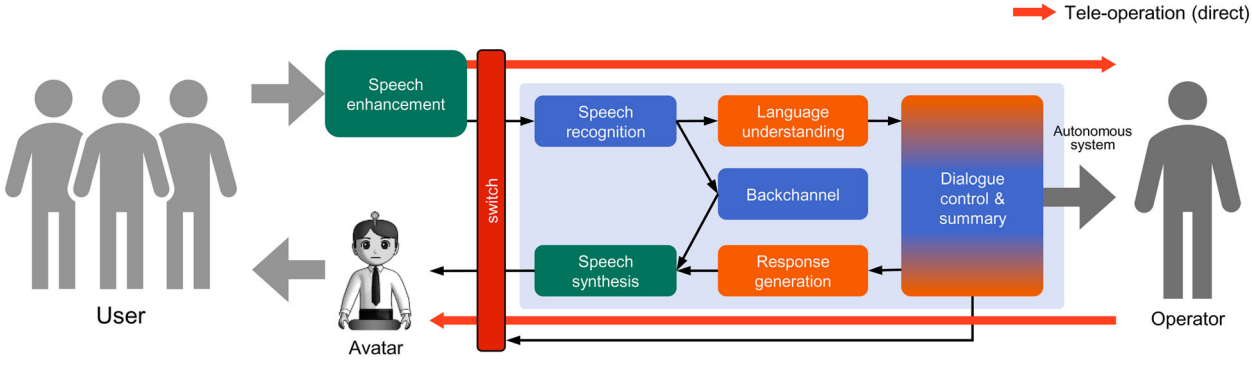

Figure 1. Proposed system architecture for semi-autonomous avatar.

Table 1. Dialogue tasks expected for semi-autonomous avatar.

	Presentation	Attentive listening	Job interview	Consulting
Major role of system	Talk	Listen	Ask	Answer
Dialogue initiative	System	User	System	Mixed
Main speaker	System	User	User	Both
Main listener	User	System	System	Both
Turn-taking	Explicit	Few	Explicit	Complicated

2.3. Technical challenges

There are many technical challenges in realizing the proposed semi-automated avatar. We need to improve autonomous dialogue systems as well as front-end speech processing to be close to the natural human level. In order to create seamless switching between the system and the operator, synthesized speech needs to match the operator's speech. This requires speech synthesis either customized for each operator or combined with voice conversion. We plan to investigate among many choices to realize high-quality voice for any individuals.

Moreover, the system must detect when it cannot cope with the user's requests, or it cannot generate an appropriate response. This is not so easy as many AI or pattern recognition systems do not know the errors by themselves. Although several attempts have been made on semi-autonomous teleoperated robots, the dialogue breakdown was conventionally detected manually or a fixed phrase such as 'That is not right' [5–7]. In the field of natural language processing, studies have been made on automatic evaluation of dialogue responses [8] and the detection of dialogue breakdown [9], but the performance is not satisfactory. When the system switches to the human operator, it should make a concise summary of the dialogue context, so the operator can promptly catch up. This requires not only a discourse summarization technique but also an effective user interface. While these technical components have been studied, they need to be extended and tailored to the proposed system.

Finally, the system should keep track of the sequence of the dialogue, including the preference of the users, for continuous improvement and better user experiences over time.

3. Parallel attentive listening system

In attentive listening, the user mainly talks while the system listens and interjects to stimulate the conversation. As regular verbal communication is important for maintaining the cognitive level and active life, attentive listening is conducted for senior care houses and local communities. As the number of volunteers is limited and their services are constrained due to COVID-19, autonomous or semi-autonomous systems are explored. We have implemented a parallel attentive listening system by integrating our base attentive listening system with dialogue monitoring and tested it in a pilot experiment.

3.1. System architecture

An example of the processing flow of the parallel attentive listening system is illustrated in Figure 2. Our attentive listening system runs on each laptop of an individual user and responds to the user's utterances. At the same time, the quality of each dialogue is monitored. Then the system switches from the autonomous system to the operator for a user who needs a human-level interlocutor to continue his/her dialogue. The operator directly talks with the user for a while to recover the dialogue, and then the switching decision will be applied back to the autonomous system. In the following implementation, we made this switching decision once every minute. Note that the operator monitors all the users sequentially for the initial minute of the dialogue.

The base attentive listening system [2] is briefly explained below. The system generates various listener responses such as backchannels, repeats, elaborating questions, assessments, generic responses as depicted in Figure 3. Backchannels are short responses such as '*Yeah*' in English and '*Un*' in Japanese. Repeats and elaborating questions are generated based on a focus word of the user utterance to express understanding by the system. Assessments are generated by a sentiment analysis to show empathy towards the user. Backchannel prediction is made for every time frame during the user's turn so that backchannels are generated even before the end of the user utterances. When the system takes the floor, one of the other responses is selected based on the selection priority order shown in Figure 3 and is uttered.

3.2. Dialogue monitoring for detecting breakdown

The current system achieves attentive listening to some extent, making 5–7 min conversations with senior people. However, it is limited compared with humans, for example, it does not show true empathy and it cannot answer complex or even simple questions. We also observed that some people stop talking when they cannot find anything to talk and the suggestions from the system are limited.

Our goal is to detect whether or not there is or will be a potential breakdown during attentive listening, based on monitoring on several modalities. We propose a basic hierarchical model with three levels representing indicators of communication breakdown, as shown in Figure 4. These are:

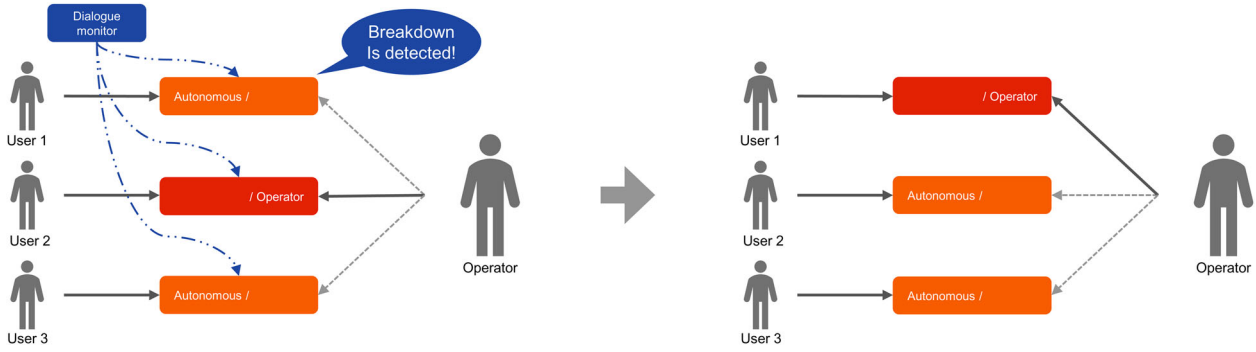

Figure 2. An example of processing flow of parallel attentive listening.

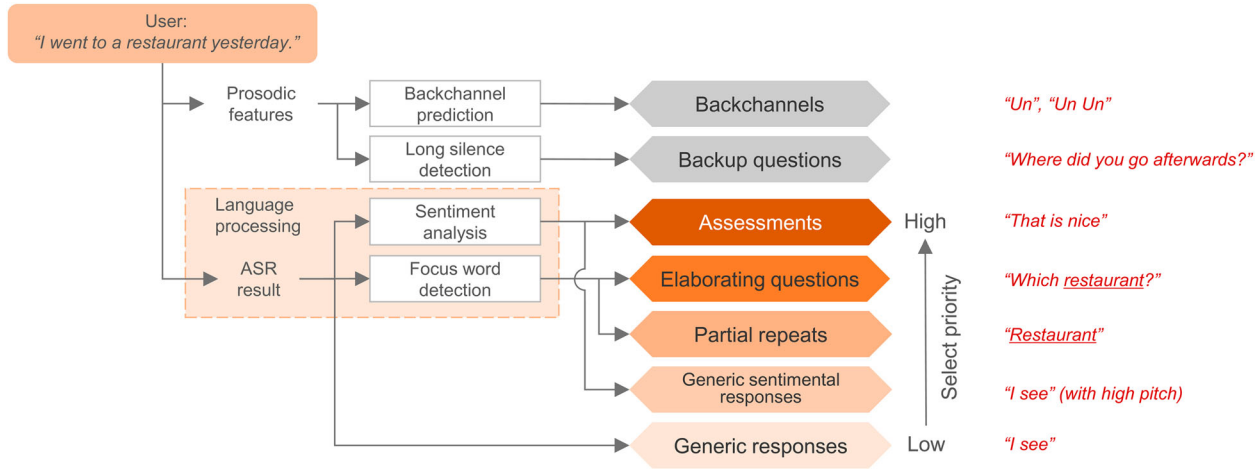

Figure 3. List of listener responses in our attentive listening system (Examples of generated responses are shown in right side in this figure. Underline means the focus word).

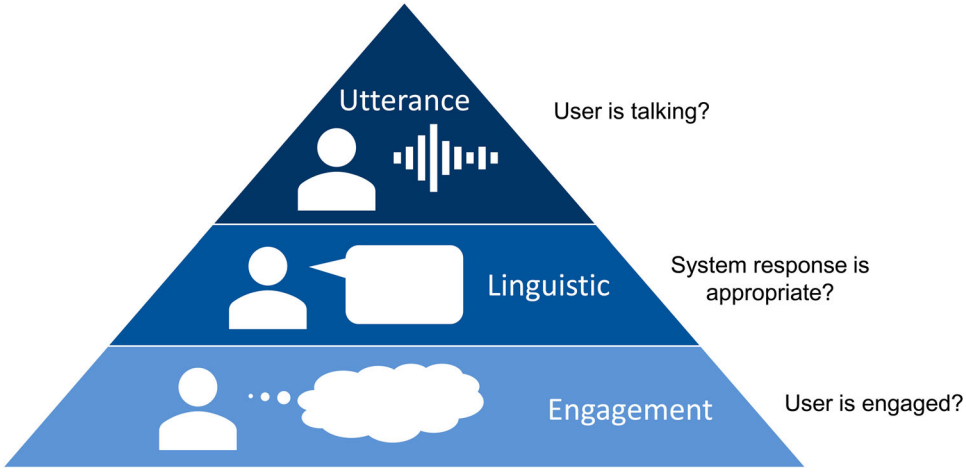

Figure 4. Levels used to determine operator switching (Higher levels are prioritized).

(1) Utterance level – the user is not talking for a certain threshold of time.
(2) Linguistic level – the generated response by the system is not appropriate.
(3) Engagement level – the user is not interested in the current conversation.

This hierarchy determines the order in which the indicators are checked.

We presume that if the user is not talking at the utterance level, then the system should switch to the human operator immediately. One cause of this may be that the user is struggling to continue the conversation, which

is reasonable after having talked for an extended period of time. In the later pilot experiment, we empirically set the threshold at 10 s. The operator could then suggest another topic or ask a question to stimulate further talk. Although breakdown detection at the utterance level is simply done with voice activity detection, the other two levels require more sophisticated techniques to decide if switching should occur.

At the linguistic level, once speech has arrived from the user and the system responds to it, the monitor evaluates the appropriateness of the response turn by turn. We can turn to a large variety of dialogue evaluation models that have been applied to linguistic input [8,10,11], which decide if the response from the system is appropriate enough. In the current prototype, we fine-tune a large-scale pre-trained model BERT [12] with appropriateness labels annotated in our previous study [2], where each system response was annotated by binary: appropriate or not. The accuracy of the model was 68.4%. If the system detects inappropriateness in more than half of system utterances during the current segment (one minute), a breakdown is detected and the autonomous system is switched to the operator.

Even if the responses by the system are correct, the user may experience boredom or disinterest in the conversation for whatever reason. This situation manifests itself at the engagement level and behaviors and utterance patterns may indicate this internal state [13]. Our system measures the engagement of the user by considering both the user and system utterances. We manually annotated the binary engagement level (engaged or not) of the user for each segment (one minute) of the above-mentioned attentive listening dialogue data [2]. By analyzing the dialogue data, we selected the feature set detectable in real-time from both user and system utterances. For example, the feature set of user utterances includes the numbers of named entities and content words (noun, verb, adjective, and adverb) in order to measure how much the user is talking substantially. Those of system utterances contain such features as the number of assessment and generic responses. Our assumption is that more generic responses (e.g. 'I see.') are less likely to keep the user engaged in the conversation. Besides these linguistic features, we are considering use of non-linguistic features such as backchannels, laughing, head nodding, and eye-gaze [14]. We trained the engagement recognition model using only linguistic features and confirmed that the recognition accuracy was 70.0%. If a switch is deemed necessary by the system at this level, the operator will be switched to the user and decide how to regain user engagement by suggesting a different topic of conversation.

Note that we use the above-mentioned models in the later pilot experiment but this is a preliminary implementation because the focus of the current paper is to show the potential and challenges of parallel attentive listening.

3.3. Pilot system

We have implemented a prototype of parallel attentive listening system to see how it works in a pilot experiment. Figure 5 is the snapshot of the system and GUI seen by

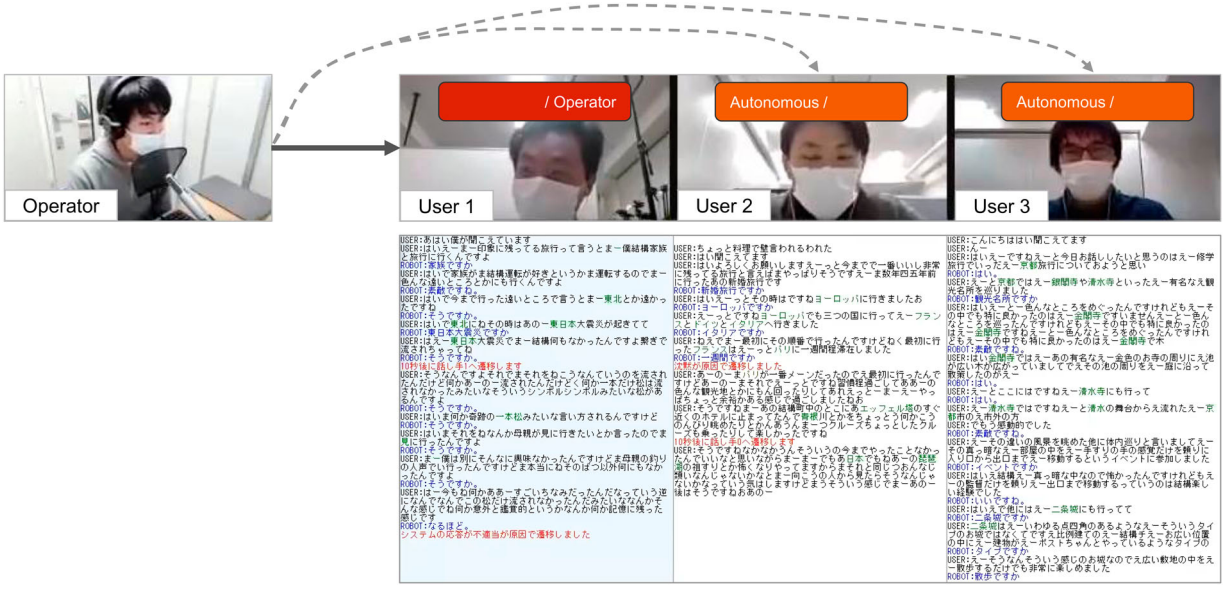

Figure 5. Snapshot of parallel attentive listening system (The red-font message shows the switch between the autonomous system and the operator and its cause. In the middle, the autonomous dialogue with User 2 was identified as a breakdown on the utterance level due to user silence and thus switched to the operator. At the end of this figure, the dialogue with User 1 was identified as a breakdown on the linguistic level due to inappropriate system responses. The operator is switched to User 1, leaving User 2).

the operator. The GUI shows the results of the automatic speech recognition and system responses. The background color represents the state of each user: blue and white mean that the user is talking with the operator and the autonomous system, respectively. To make the operator easily and efficiently understand the context of each dialogue, named entity words such as names of places and persons are highlighted. When the system has made the switching decision, advance notice is displayed 10 s before the switching. Then the cause of switching (the type of breakdown that led to the decision to intervene) is also displayed. As an interface between the system and each user, we used static images so that each user could see which type of the system (operator or autonomous) he/she is talking with.

We then tested this system with three users simultaneously and compared it with the baseline setting where the users talked with the autonomous system without any human intervention (*fully autonomous*). The subjects were 18 university students and equally divided into two conditions: *semi-autonomous* (proposed) and *fully autonomous* (baseline). It was confirmed that the operator could monitor the parallel attentive listening with the proposed pilot system and could intervene in a user who seemed to be difficult to talk with the autonomous system. We also asked them to evaluate each system by evaluation metrics designed by referring to those used in our previous experiment with the fully autonomous system [2]. The result suggests that the pilot system obtained higher scores than the fully autonomous system on items about such as *understanding*, *smoothness*, and *satisfaction*. Based this pilot study, we will improve the proposed system and conduct further experiments to fill the gap between the current semi-autonomous system and human dialogue.

4. Ongoing work

In this paper, we have described multiple semi-autonomous conversational avatars which are simultaneously tele-operated by a single human. This solution mitigates not only physical constraints but also time constraints with the ability to conduct tasks with multiple users, which will greatly increase efficiency. We designed and implemented the framework for the task of attentive listening which allows the system to know when to switch to the operator.

There are several areas in which we plan to extend this work – the number and type of avatars, the range of tasks, and the number of modalities. Our basic prototype had the operator simultaneously managing three avatars. The goal is to obviously extend this to as many avatars as possible while keeping the operator's cognitive load at a manageable level. If the task is not so complex and a strong AI can be built for it, then a large number of avatars can be handled since the operator only has to intervene for a small number of edge cases. In this pilot system, the avatars themselves were static images and responded using voice alone. It will be implemented with many types of robots and virtual agents which can also communicate non-verbally. Our goal is to make this system independent of any particular avatar, so that novice operators may freely control their own robots through the use of APIs, without modifying the underlying logic. Our prototype focused on the task of attentive listening. As we described earlier, we have several other potential applications with different requirements and dialogue models. These types of tasks, for example consulting, may also need specially trained operators, which should be considered. The input modalities of the user must also be extended. Our prototype used just one audio modality with speech recognition, but the extension of this is streamed video which can be used as input by the system for better monitoring.

The outcomes of this research will enhance the society through the new normal – the ability to efficiently conduct activities which require social interaction even while being physically distant, and also being able to interface with multiple remote users simultaneously. This type of solution will be suitable even after the pandemic has ended.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

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

Notes on contributors

Tatsuya Kawahara received B.E. in 1987, M.E. in 1989, and Ph.D. in 1995, all in information science, from Kyoto University, Kyoto, Japan. From 1995 to 1996, he was a Visiting Researcher at Bell Laboratories, Murray Hill, NJ, USA. Currently, he is a Professor and the Dean of the Graduate School of Informatics, Kyoto University. He was also an Invited Researcher at ATR and NICT. He has published more than 400 academic papers on speech recognition, spoken language processing, and spoken dialogue systems. He has been conducting several projects including speech recognition software Julius, the automatic transcription system deployed in the Japanese Parliament (Diet), and the autonomous android ERICA. He received the Commendation for Science and Technology by the Minister of Education, Culture, Sports, Science and Technology (MEXT) in 2012. From 2003 to 2006, he was a member

of IEEE SPS Speech Technical Committee. He was a General Chair of IEEE Automatic Speech Recognition and Understanding workshop (ASRU 2007). He also served as a Tutorial Chair of INTERSPEECH 2010 and a Local Arrangement Chair of ICASSP 2012. He was an editorial board member of Elsevier Journal of Computer Speech and Language and IEEE/ACM Transactions on Audio, Speech, and Language Processing. He is the Editor-in-Chief of APSIPA Transactions on Signal and Information Processing. He is a board member of APSIPA and ISCA, and a Fellow of IEEE.

Naoyuki Muramatsu received B.E. in 2021 from the Undergraduate School of Informatics and Mathematical Science in Kyoto University, Kyoto, Japan. His research interest includes parallel conversation by semi-autonomous spoken dialogue systems.

Kenta Yamamoto received M.S. in 2020 and is currently pursuing a Ph.D. degree at the Graduate School of Informatics in Kyoto University, Kyoto, Japan. He is also a JSPS Research Fellow (DC1). His research interests include spoken dialogue systems (SDSs) and character expression for SDSs.

Divesh Lala received his Ph.D. in 2015 from the Graduate School of Informatics in Kyoto University, Kyoto, Japan. Currently, he is a specially appointed researcher at the same institution. His research interests include human-agent interaction and multimodal signal processing.

Koji Inoue received Ph.D. in 2018 from the Graduate School of Informatics in Kyoto University, Kyoto, Japan. He was a JSPS Research Fellow (DC1) from 2015 to 2018. Currently, he is an Assistant Professor of the Graduate School of Informatics, Kyoto University. His research interests include spoken dialogue systems, speech signal processing, and human-robot interaction.

References

[1] Kawahara T. Spoken dialogue system for a human-like conversational robot ERICA. In: International Workshop on Spoken Dialog System Technology (IWSDS), Singapore; 2018.

[2] Inoue K, Lala D, Yamamoto K, et al. An attentive listening system with android ERICA: comparison of autonomous and WOZ interactions. In: Sigdial Meeting on Discourse and Dialogue (SIGDIAL); 2020. p. 118–127 [Online].

[3] Inoue K, Lala D, Yamamoto K, et al. Engagement-based adaptive behaviors for laboratory guide in human-robot dialogue. In: International Workshop on Spoken Dialog System Technology (IWSDS), Siracusa; 2019.

[4] Inoue K, Hara K, Lala D, et al. Job interviewer android with elaborate follow-up question generation. In: International Conference on Multimodal Interaction (ICMI); 2020. p. 324–332 [Online].

[5] Glas DF, Kanda T, Ishiguro H, et al. Simultaneous teleoperation of multiple social robots. In: International Conference on Human-Robot Interaction (HRI), Amsterdam; 2008. p. 311–318.

[6] Shiomi M, Sakamoto D, Kanda T, et al. A semi-autonomous communication robot –a field trial at a train station–. In: International Conference on Human-Robot Interaction (HRI), Amsterdam; 2008. p. 303–310.

[7] Kanda T, Shiomi M, Miyashita Z, et al. A communication robot in a shopping mall. IEEE Trans Robot. 2010;26(5):897–913.

[8] Lowe R, Noseworthy M, Serban IV, et al. Towards an automatic Turing test: Learning to evaluate dialogue responses. In: Annual Meeting of the Association for Computational Linguistics (ACL), Vancouver; 2017. p. 1116–1126.

[9] Higashinaka R, D'Haro LF, Shawar BA, et al. Overview of the dialogue breakdown detection challenge 4. In: International Workshop on Spoken Dialog System Technology (IWSDS), Siracusa; 2019.

[10] Tao C, Mou L, Zhao D, et al. RUBER: An unsupervised method for automatic evaluation of open-domain dialog systems. In: AAAI Conference on Artificial Intelligence (AAAI), New Orleans; 2018. p. 722–729.

[11] Zhao T, Lala D, Kawahara T. Designing precise and robust dialogue response evaluators. In: Annual Meeting of the Association for Computational Linguistics (ACL); 2020. p. 26–33 [Online].

[12] Devlin J, Chang MW, Lee K, et al. BERT: pre-training of deep bidirectional transformers for language understanding. In: North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT), Minneapolis; 2019. p. 4171–4186.

[13] Oertel C, Castellano G, Chetouani M, et al. Engagement in human-agent interaction: an overview. Front Robot AI. 2020;7:92.

[14] Inoue K, Lala D, Takanashi K, et al. Engagement recognition by a latent character model based on multimodal listener behaviors in spoken dialogue. APSIPA Trans Signal Inf Process. 2018;7:e9.