Development and evaluation of a semi-autonomous parallel attentive listening system

Divesh Lala, Koji Inoue, Haruki Kawai, Zi Haur Pang, Mikey Elmers and Tatsuya Kawahara
Graduate School of Informatics, Kyoto University, Japan
E-mail: [lala][inoue][pang][elmers][kawahara]@sap.ist.i.kyoto-u.ac.jp

***Abstract*—Open domain spoken dialogue systems are still not at the level of human performance and understanding. In this work we propose that the a remote human operator who can take over the conversation from an autonomous system would improve the quality of an attentive listening conversation. Furthermore this operator could also manage multiple users simultaneously. We describe and implement this as a semi-autonomous parallel system. Features of this system are detection of disengagement to let the operator know when to intervene, summarization of conversations using ChatGPT to allow the operator to manage multiple users, and conversion of the operator's voice to the agent's to make intervention less abrupt. We conduct an experiment to compare this system to a fully autonomous system and find that it improves performance for enjoyment and empathy.**

I. Introduction

Through the use of large language models (LLMs) and improved automatic speech recognition (ASR) technology, spoken dialogue chatting systems have advanced to become more human-like. However, there is still a clear gap between human-human and human-agent conversation.

Most commercial spoken dialogue systems, such as smart speakers, are primarily task-based. However chatting systems are yet to reach this level of wide usage. This is arguably because such open-domain systems are more complex and require servicing a variety of individual users. LLM dialogue systems such as ChatGPT are now becoming popular, although using these in natural spoken conversation still requires real-time conversation functions such as human-like turn-taking and backchannelling.

LLM-based approaches are also not immune to making inappropriate responses (hallucinations) that arise as a result of either inaccurate ASR results or the model itself. Many dialogue systems do not understand when they make errors and so continue without acknowledgement. On the other hand, a human can recognize and rectify the situation through an apology or social behavior to express this to the user.

Recently there has been much work investigating how LLMs can show this type of empathy in a dialogue system, with some degree of success [1]–[3]. However LLMs and artificial systems can only show this in a synthetic way without being able to feel emotions. Furthermore, empathetic expression is still difficult to replicate through text-to-speech technologies.

To bridge this gap we propose to combine the conversation skills of humans with an autonomous system to make a semi-autonomous system. In this system a remote human operator can smoothly "take over" the conversation from an agent and provide human-level intelligence and empathy at certain times. This approach has been implemented in previous work for service agents [4], [5], but is redundant for open-domain conversation with only one user.

We therefore consider a schema where multiple users interact with separate spoken chat systems simultaneously. In this case, the semi-autonomous function is used by the operator to freely switch between each human-agent conversation and intervene where necessary. We define this setup with independent users as a parallel system. Fig. 1 shows a general outline of the semi-autonomous parallel approach.

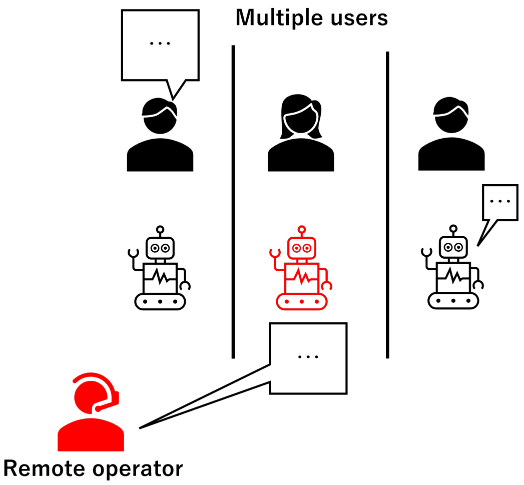


Fig. 1. Overview of the semi-autonomous parallel approach. A remote operator monitors several independent users and can intervene in a conversation, taking over from the autonomous agent.

This system needs to address several challenges such as deciding when the operator should intervene in the conversation and how to allow the operator to immediately understand the conversation in real-time. In this work we focus on attentive listening, an open-domain dialogue system. The semi-autonomous system can be implemented in areas where online agents are used for the benefit of multiple end-users, such as guidance, interviews and teaching.

We expect that a semi-autonomous system can outperform an autonomous system in attentive listening by exploiting a human's conversation skill. Our goal in this work is to implement the system and compare it against a fully autonomous dialogue system. The work in this study is conducted in the

Japanese language.

II. Related work

Fully tele-operated robots have been implemented in various service and therapeutic tasks, with either the operator communicating directly as a human being [6], [7] or attempting to imitate a robot [8], [9]. In these cases one operator is responsible for a single agent and not a parallel implementation.

The development of semi-autonomous systems, where a human can intervene on behalf of an agent, has been achieved with customer service-related tasks [4], [5] and chatting [10], although these are for text-based chatting and not spoken dialogue. A recent work implemented a semi-autonomous system in spoken dialogue but only in a dyadic context [11].

The parallel architecture has also been implemented in a semi-autonomous system with multiple mobile robots [12], although the operator did not directly speak to users, instead they communicated through a limited fixed set of responses. Additionally, [13] previously implemented a semi-autonomous parallel system for a question-answering task. However chatting has much different requirements in terms of handover. To our knowledge this work presents the first semi-autonomous parallel system for an open-domain spoken dialogue task.

III. Autonomous attentive listening system

The dialogue system used in this experiment is an attentive listening system where the agent acts as a listener towards the user's talk, such as a recent enjoyable experience. The objective is for the user to speak to the system and for the conversation to slowly encompass a range of topics. However if the user becomes disengaged due to inappropriate or monotonous responses from the system they will find it difficult to speak for a lengthy period of time. Furthermore, since the dialogue system does not take initiative from the user it must produce meaningful questions and empathetic responses which stimulate more conversation.

This system uses a state-of-the-art transformer-based Japanese ASR model [14]. Turn-taking and backchannelling are handled by our previously implemented models REF. Backchannels in response to an emotional state (e.g. "ah!", "oh!") are also generated as the system detects user sentiment.

The system has four types of responses - sentiments, elaborating questions, repeated responses and generic responses, summarized in Fig. 2.

Sentiments are a reaction containing some type of emotion (e.g. "Wow, that's great" (*ii desu ne*) or "That's a shame" (*zannen desu ne*)) and are prioritized over the other types of response. We use a simple lexical model to detect a sentiment.

A. Focus word detection

Our approach requires detecting the focus of the user's talk which we implemented in previous work [15]. For example, if the user says "Today I went to a restaurant and ate pasta", the focus word *pasta* would be detected.

The focus word is also used for repeated responses, which are simply the focus word with *desu ka* added to it. In Japanese, these types of responses act like affirmations or backchannels rather than direct questions. To ensure a variety of responses, the system will maintain a balance between elaborating questions and repeated responses.

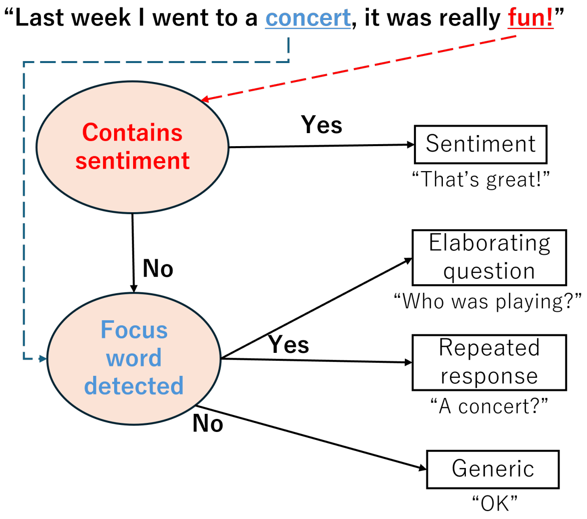


Fig. 2. Overview of the response types used in the attentive listening system.

Generic responses are used when no focus word can be detected and are simple statements such as "I see" (*sou desune*) or OK (*hai*). Our aim is to keep these to a minimum.

B. Response generation with T5 model

If a focus word is detected we generate an elaborating question using a specialized deep learning model distilled from ChatGPT. We first extracted data from RealPersonaChat [16], a corpus of 14,000 dialogues. The data was statements which preceded a question. These were then used with the following prompt to ChatGPT (translated from Japanese):

> *"Please respond to the statements with elaborating questions to elicit more conversation. The purpose of the elaborating questions is to get important information about the 5W1Hs: "why", "what", "where", "who", "when" and "how"."*

This prompt produces responses which are more ideal for attentive listening than the longer utterances used by naive ChatGPT. An example of the different responses is:

User input: I ate lunch outside but I had a bad feeling so I came back.

ChatGPT (naive) That is unfortunate. Was something not to your liking? Was there something wrong with the location or the food?

ChatGPT (prompt) Why did you have a bad feeling?

This resulted in training data consisting of around 16,000 statement and response pairs used to fine-tune a T5 model [17]. The model is smaller than a typical LLM (around 220 million parameters), but produces attentive listening-type dialogue. The resulting model is lightweight, produces responses suitable for attentive listening and can be used with a CPU.

The system also uses elaborating questions in situations when the conversation seems to be stalling. When there is a silence from the user of greater than 5 seconds an elaborating question is uttered using a history of the dialogue. Elaborating questions are also generated when three or more generic responses are generated consecutively by the system in order to preserve response variety.

C. Agent interface

The agent we use in this work is Gene, based on the MMDAgent platform [18], shown in Fig. 3. Gene can use a variety of text-to-speech (TTS) systems. In this case we use a female voice. Gene has lip syncing capabilities and conveys emotional states through facial expressions. When sentiment responses or emotional backchannels are uttered by Gene, they are combined with the appropriate expression.

Fig. 3. Gene, the agent used in this work, with a sad expression.

D. Limitations of autonomous attentive listening

Although this system is robust there are limitations which could be improved by human intervention. The elaborating questions are generally appropriate but quite limited. If a focus word cannot be found the system reverts to generic responses which do not stimulate more talk and it may become difficult for the user to speak fluently. We also showed that users still found autonomous attentive listening to be lacking in empathy and understanding compared to a tele-operated system [15].

These issues can be alleviated with human intervention. An operator can quickly come up with more probing questions to promote talk, quickly recognize dialogue breakdowns, and assist the user in a similar manner. This motivates our semi-autonomous approach.

IV. Semi-autonomous system

We now describe the main aspects of the semi-autonomous system which are disengagement detection (to determine when the operator should intervene) and dialogue monitoring (so the operator quickly understands the state of the dialogue).

A. Disengagement detection

The first technical challenge is detecting when the system needs human intervention because of dialogue breakdown or disinterest. Although this has been widely studied [19], it is more difficult in attentive listening since the user may try to keep talking even in an uninteresting conversation. In this work we estimate if the user may require intervention by continuously analyzing aspects of the attentive listening dialogue itself.

To approach this problem, we use the set of heuristics described in [11]. These are based on the assumptions that a "good" attentive listening dialogue involves the user talking continuously.

- User is silent for more than five seconds
- User has had two short turns (less than 20 characters) in a row
- System has not uttered anything except generic responses for three turns in a row
- System has not uttered a sentiment response in the last five turns

The first two heuristics estimate the user state in the dialogue. The final two heuristics relate to the performance of the system. The operator is provided information when one or more of the conditions are met, in the form of recommendations to intervene. The operator is not required to act upon these recommendations but they are useful for understanding what type of dialogue intervention may be needed to help maintain the conversation.

B. Dialogue monitoring

The operator must be able to monitor multiple conversations without too much cognitive load. If the operator has to read ASR results it will take too long to "catch up" to the conversation. For a parallel system this is even more critical as the operator has to deal with multiple users without delay.

To approach this problem we use dialogue summarization through ChatGPT as proposed in previous work [10]. The performance of ChatGPT in summarization tasks is comparable to humans [20] and output can be verified in real-time by the operator. We use few-shot prompting for ChatGPT using the following instruction (translated from Japanese):

"Below is part of a conversation between A and B. Summarize A's entire talk, their emotional state and provide simple short questions to ask A. Ignore any unnatural Japanese in the dialogue. Also please ignore B's dialogue. The output should be less than 100 Japanese characters. There are some examples below."

The prompt includes several examples and the most recent dialogue of the conversation from ASR results. The user and agent correspond to speaker A and B respectively. The output format from ChatGPT is in three parts - a short textual summary of the conversation, an estimation of the emotional state of the user, and questions that could be asked to the user. The intention is that the operator uses these to formulate more in-depth questions with appropriate backchannels and sentiment reactions. Summarization is executed when the operator clicks on a user's video panel.

C. Seamless switching

Another major challenge is to match the voice of the operator and the agent's TTS to prevent an abrupt change when control is switched between them. We use voice conversion [21] to change the operator's voice to one which matches the agent's TTS in real-time. This allows the operator to speak as the agent while preserving the nuances of human speech.

V. Operator interface

Fig. 4 shows the operator interface. The operator can view ASR results and system dialogue in real-time. They select the conversation they would like to listen to by clicking on the video panel. The agent is not displayed. To interact with the user, the operator clicks a microphone button corresponding to the conversation in which they wish to intervene. This button is disabled when the agent is speaking to prevent interference between the operator and agent's voice. During an intervention the operator may also control particular facial expressions of the agent by clicking the corresponding emoji button.

We evaluated the interface in an initial feasibility study with two operators over 10 sessions. We used a 5-point Likert scale with six items and found that all items averaged at least 3. The usability of the interface (4.4) and the ability to talk spontaneously to the users (4.3) were rated the highest.

VI. Experiment methodology

We hypothesize that a semi-autonomous system with one user is more effective than one with multiple users and that summarization improves the system. We conducted experiments under four different conditions:

AUTO The autonomous attentive listening system.

DYADIC A semi-autonomous system where the operator only handled one user with no summarization.

SEMI-BASE A semi-autonomous parallel system where the operator handled three users with no summarization.

SEMI-FULL A semi-autonomous parallel system where the operator handled three users using summarization.

The **DYADIC** condition was implemented in [11] with the same attentive listening system and experimental procedure.

We first conducted an experiment with the **SEMI-BASE** system. There were six sessions with three simultaneous participants. We then conducted a within-subjects experiment with 11 participants who interacted with both the **SEMI-FULL** and **AUTO** systems. For the **SEMI-FULL** system we conducted four sessions. One session had two participants so we used a "dummy" participant, to ensure that the operator would still manage three users. The order of conditions was randomized.

Participants were given an explanation of attentive listening and time to consider what they would talk about. The motivation topic was something interesting that they had experienced recently. We informed participants that their conversations would be listened to by a remote operator who could intervene on behalf of the agent, although they would not know when it would happen. We used two different operators for these conditions. Each session was around eight minutes in duration. Experiments were conducted with both participants and operators in sound-proof cubicles in the same room and the operator cubicle was separated from the participants so they could not see each other.

Operators in the semi-autonomous conditions were instructed to intervene in conversations by considering information provided by the interface, although they would make the final intervention decision themselves. They were instructed to only use short questions and statements suitable for attentive listening. Note there is no difference between the **SEMI-BASE** and **SEMI-FULL** systems from the user perspective, only the information the operator receives.

VII. Results

After each interaction participants answered a 7-point Likert survey consisting of 19 items based on those found in previous work [15]. We grouped these items to construct measures and used them as the basis of the statistical analysis.

Likert scale results are shown in Fig. 5, taking the average item for each measure. We could not outperform the **DYADIC** condition, which had the highest ratings across all measures, even without summarization. However **SEMI-FULL** is close to the level of the **DYADIC** system, showing that summarization is beneficial when handling multiple users. **SEMI-BASE**, which did not provide any summarization, performs worse than **SEMI-FULL** on all measures.

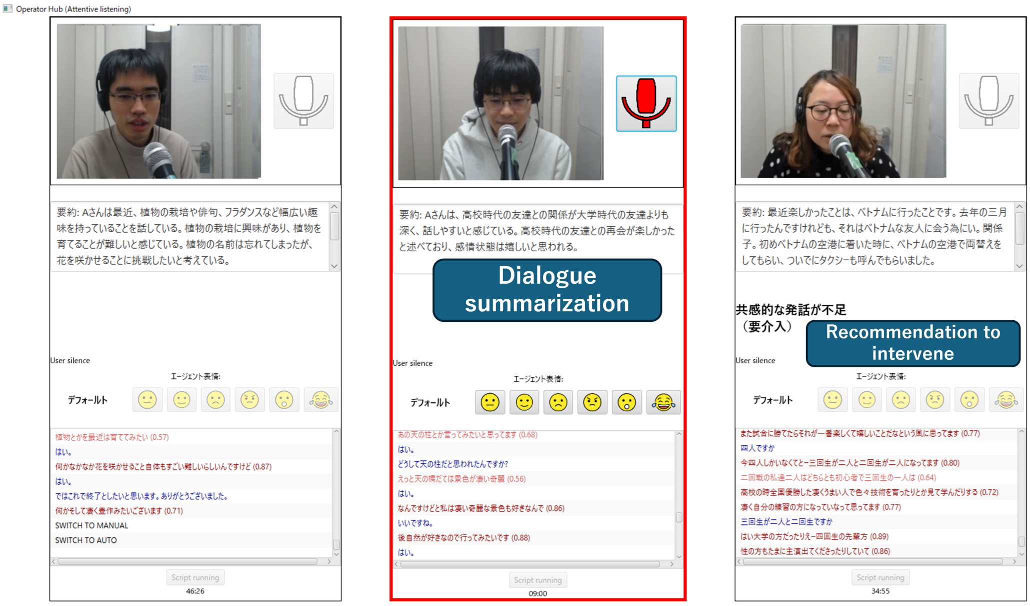
The **SEMI-FULL** system also outperformed **AUTO**, particularly for empathy and enjoyment. Paired t-test results for these measures showed p-values of 0.07 and 0.08 respectively. These are also the only two measures where the medians of **SEMI-BASE** are higher than that of **AUTO**.

On the other hand, the **SEMI-BASE** system was worse than the **AUTO** system in terms of naturalness and timing. From this result we propose that if the operator cannot provide a quality intervention then the users perceive it as upsetting the flow of the conversation. On the other hand the **SEMI-BASE** system was better in terms of enjoyment and empathy but there is a large amount of variation between users. The average number of interventions per user was around 3, but the number of interventions did not correlate with any measures.

VIII. Discussion

Results of our study indicated that the semi-autonomous approach improved enjoyment and empathy but naturalness and timing were perceived more negatively. One reason for this could be that the items which make up naturalness and timing are arguably more to do with conversational processes as opposed to how users feel about the agent.

The effect of human intervention changes with both the number of simultaneous users and the existence of summarization. Although intervention can be done successfully in a



要約: Aさんは最近、植物の栽培や俳句、フラダンスなど幅広い趣味を持っていることを話している。植物の栽培に興味があり、植物を育てることが難しいと感じている。植物の名前は忘れてしまったが、花を咲かせることに挑戦したいと考えている。

植物とかを最近は育ててみたい (0.57)
はい。
何かなかなか花を咲かせること自体もすごい難しいらしいんですけど (0.87)
はい。
ではこれで終了としたいと思います。ありがとうございました。
何かそして凄く豊作みたいございます (0.71)
SWITCH TO MANUAL
SWITCH TO AUTO

要約: Aさんは、高校時代の友達との関係が大学時代の友達よりも深く、話しやすいと感じている。高校時代の友達との再会が楽しかったと述べており、感情状態は嬉しいと思われる。

あの天の柱とか言ってみたいと思ってます (0.68)
はい。
どうして天の柱だと思われたんですか?
えっと天の橋だては景色が凄い奇麗 (0.56)
はい。
なんですけどと私は凄い奇麗な景色も好きなんで (0.86)
いいですね。
後自然が好きなので行ってみたいです (0.88)
はい。

要約: 最近楽しかったことは、ベトナムに行ったことです。去年の三月に行ったんですけれども、それはベトナムな友人に会う為にい。関係子。初めベトナムの空港に着いた時に、ベトナムの空港で両替えをしてもらい、ついでにタクシーも呼んでもらいました。

共感的な発話が不足
（要介入）

また試合に勝てたらそれが一番楽しくて嬉しいことだなという風に思ってます (0.77)
四人ですか
今四人しかいなくてとー三回生が二人と二回生が二人になってます (0.80)
二回戦の私達二人はどちらとも初心者で三回生の一人は (0.64)
高校の時全国優勝した凄くうまい人で色々技術を育ったりとか見て学んだりする (0.72)
凄く自分の練習の方になっていなって思ってます (0.77)
三回生が二人と二回生ですか
はい大学の方だったりえー四回生の先輩方 (0.89)
性の方もたまに主演出てくださったりしていて (0.86)

Fig. 4. User interface of the remote operator. The currently selected user is in the middle panel and the operator can speak to them directly. Intervention recommendations and dialogue summarization are labeled.

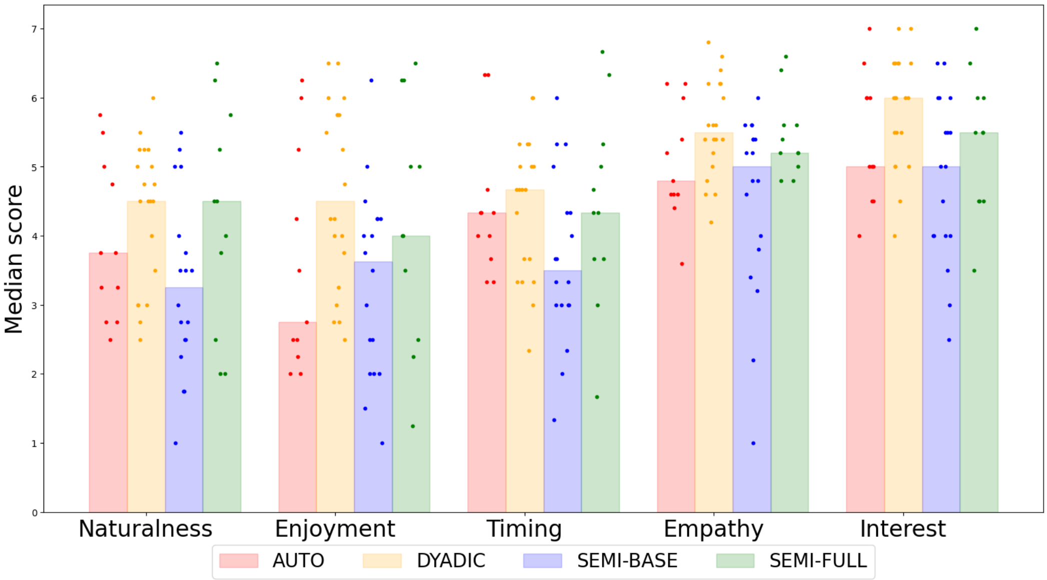


Fig. 5. Results of experiment. Bars indicate median values.

dyaduc interaction, for multiple users summarization is necessary to quickly understand the state of the dialogue. Without summarization the *quality* of the intervention is degraded because they cannot successfully stimulate the conversation.

Other strategies which can assist with quality interventions should be considered. Our eventual aim is to at least match the performance of the **DYADIC** system, which we used as a comparative baseline. This would show that one operator can service the needs of multiple people simultaneously at the same level as a single user. Limitations of this work were the relatively low number of participants in the study and the quality of the voice conversion.

IX. Conclusion

In this work we introduced a semi-autonomous parallel system for an attentive listening spoken dialogue system. A remote operator intervenes and takes control of a human-agent conversation to help maintain user interest. The operator can manage several of these dialogues simultaneously through our system, by using disengagement detection and dialogue summarization. Our user study showed that it could improve user perception of enjoyment and empathy and that dialogue summarization was improved the performance of the system.

Acknowledgment

This work was supported by JST Moonshot R&D Goal 1 Avatar Symbiotic Society Project (JPMJPS2011).

References

[1] J. W. Ayers, A. Poliak, M. Dredze, *et al.*, "Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum," *JAMA internal medicine*, vol. 183, no. 6, pp. 589–596, 2023.

[2] Z. Elyoseph, D. Hadar-Shoval, K. Asraf, and M. Lvovsky, "Chatgpt outperforms humans in emotional awareness evaluations," *Frontiers in Psychology*, vol. 14, 2023.

[3] Y. Fu, K. Inoue, D. Lala, K. Yamamoto, C. Chu, and T. Kawahara, "Dual variational generative model and auxiliary retrieval for empathetic response generation by conversational robot," *Advanced Robotics*, vol. 37, no. 21, pp. 1406–1418, 2023.

[4] M. Poser, T. Hackbarth, and E. A. Bittner, "Don't throw it over the fence! Toward effective handover from conversational agents to service employees," in *International Conference on Human-Computer Interaction*, Springer, 2022, pp. 531–545.

[5] M. Poser, S. Singh, and E. Bittner, "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*, 2021, p. 1181.

[6] K. Kuwamura, S. Nishio, and S. Sato, "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*, vol. 7, p. 1066, 2016.

[7] L. Chen, H Sumioka, L Ke, M Shiomi, and L Chen, "Effects of teleoperated humanoid robot application in older adults with neurocognitive disorders in taiwan: A report of three cases," *Aging Medicine and Healthcare*, vol. 11, pp. 67–71, 2 2020.

[8] J. Baba, S. Sichao, J. Nakanishi, *et al.*, "Teleoperated robot acting autonomous for better customer satisfaction," in *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*, Association for Computing Machinery, 2020, pp. 1–8.

[9] S. Song, J. Baba, J. Nakanishi, Y. Yoshikawa, and H. Ishiguro, "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*, Association for Computing Machinery, 2021, pp. 1–6.

[10] S. Yamashita and R. Higashinaka, "Data collection for empirically determining the necessary information for smooth handover in dialogue," in *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, Marseille, France: European Language Resources Association, 2022, pp. 4060–4068.

[11] H. Kawai, D. Lala, K. Inoue, K. Ochi, and T. Kawahara, "Evaluation of a semi-autonomous attentive listening system with takeover prompting," *arXiv preprint arXiv:2402.14863*, 2024.

[12] D. F. Glas, T. Kanda, H. Ishiguro, and N. Hagita, "Teleoperation of multiple social robots," *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans*, vol. 42, no. 3, pp. 530–544, 2012.

[13] Y. Muraki, H. Kawai, K. Yamamoto, K. Inoue, D. Lala, and T. Kawahara, *Semi-autonomous guide agents with simultaneous handling of multiple users*, 2023.

[14] M. Mimura, K. Inoue, T. Kawahara, T. Nakamura, and H. Saruwatari, "Construction of a corpus of spoken Japanese under real-world conditions and its use in speech recognition benchmarks," Tech. Rep. 12, 2023, (in Japanese).

[15] K. Inoue, D. Lala, K. Yamamoto, S. Nakamura, K. Takanashi, and T. Kawahara, "An attentive listening system with android ERICA: Comparison of autonomous and WOZ interactions," in *Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, O. Pietquin, S. Muresan, V. Chen, *et al.*, Eds., Association for Computational Linguistics, 2020, pp. 118–127.

[16] S. Yamashita, K. Inoue, A. Guo, S. Mochizuki, T. Kawahara, and R. Higashinaka, "RealPersonaChat: A realistic persona chat corpus with interlocutors' own personalities," in *Proceedings of the 37th Pacific Asia Conference on Language, Information and Computation*, C.-R. Huang, Y. Harada, J.-B. Kim, *et al.*, Eds., Dec. 2023, pp. 852–861.

[17] C. Raffel, N. Shazeer, A. Roberts, *et al.*, "Exploring the limits of transfer learning with a unified text-to-text transformer," *Journal of Machine Learning Research*, vol. 21, no. 140, pp. 1–67, 2020.

[18] A. Lee, *MMDAgent-EX*, version 1.0.0, 2023. [Online]. Available: https://github.com/mmdagent-ex/MMDAgent-EX.

[19] R. Higashinaka, L. F. D'Haro, B. Abu Shawar, *et al.*, "Overview of the dialogue breakdown detection challenge 4," in *Increasing Naturalness and Flexibility in Spoken Dialogue Interaction: 10th International Workshop on Spoken Dialogue Systems*, Springer, 2021, pp. 403–417.

[20] J. Wang, Y. Liang, F. Meng, *et al.*, "Is chatgpt a good nlg evaluator? a preliminary study," *arXiv preprint arXiv:2303.04048*, 2023.

[21] J. Kim, J. Kong, and J. Son, "Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech," in *Proceedings of the 38th International Conference on Machine Learning*, M. Meila and T. Zhang, Eds., ser. Proceedings of Machine Learning Research, vol. 139, 2021, pp. 5530–5540.