Effective Recovery Strategies in Conversations with Older Adults

Shaul Ashkenazi[1], Bonnie Webber[2], and Maria Wolters[2]

***Abstract*— Misunderstandings happen both in interactions between humans and in interactions between humans and voice assistants. Successful voice assistants know how to recover from such misunderstandings gracefully. We compared the effectiveness of two recovery strategies, AskRepeat (request user to repeat the sentence) and ReprompThenSay (repetition of prompt, followed by instruction) for younger and older users. The strategies were tested with 26 participants, 13 younger (aged 22–29) and 13 older (aged 66–81). Overall, users recovered successfully from problems they encountered with the system. Older and younger users performed equally well, and we found that ReprompThenSay was more effective for both age groups. Older users encountered more issues when using the system and were more likely to be annoyed with it, but found it as likable and habitable as younger users. We conclude that recovery strategies may need to be adapted to specific challenges and expectations instead of age.**

I. Introduction

With the advent of Large Language Models (LLMs), smart speakers such as Amazon Alexa or Google Home, and phone-based voice interfaces like Apple's Siri, more people use conversational user interfaces than ever before. In verbal interactions between users and voice assistants, just as in verbal interactions between humans, the voice assistant may not always understand what the user says. For the purpose of this paper, following McRoy [1], we define misunderstandings as cases where the user's speech was not recognised correctly, and non-understandings as cases where the user's utterance could not be interpreted or did not make sense given the current stage of the conversation.

In this paper, we focus on non-understandings, where the system is aware of a problem. Following [2], we call the system's attempt to elicit new, interpretable data in response to the problem *recovery*. While recovery strategies have been tested extensively with younger users (e.g., [3]–[5]), and there is a substantial body of research on adapting voice assistants to older users (e.g., [6]–[8]), there is relatively little work on effective recovery strategies for older users. Older users may have different mental models of how voice assistants work; at the same time, they stand to benefit more from assistants that are always available.

In an extensive comparative study of recovery strategies for a room booking system, Bohus and Rudnicky [2] found that the most successful strategy was to move on or change the direction of the dialogue, followed by more or less terse help messages and reprompting. Asking the user to repeat what they said was far less successful. Examples of these recovery strategies are listed in Table I.

TABLE I
Selected Recovery Strategies (Bohus and Rudnicky [2])

Recovery Strategy	Example
AskRepeat	*Can you please repeat that?*
Reprompt	[system repeats the previous prompt]
MoveOn	*Sorry, I didn't catch that. One choice would be Wean Hall 7220. This room can accommodate 20 people and has a whiteboard and a projector. Would you like a reservation for this room?*
TerseYouCanSay	*Sorry, I didn't catch that. You can say 'I want a small room' or 'I want a large room'. If the size of the room doesn't matter to you, just say 'I don't care'.*

Since speech does not require operating a keyboard or a touch screen, it should theoretically be a particularly intuitive interaction modality for older adults. However, older adults also face many age-specific issues. For example, ASR systems still have higher word error rates for older adults [9]. Some older adults may also treat voice assistants more as conversation partners than as computers that process commands [10].

Despite those age-related issues, there is no typical "older adult behaviour", and older people these days are far more likely to have interacted with a computer during their life than older adults 15–20 years ago. Therefore, instead of designing voice assistants that specifically cater to the needs of older people, we should make existing systems robust enough to work well for the majority of all adults. Good recovery strategies are key to robustness.

In this study, we assessed to what extent these recovery strategies could support recovery from non-understandings with older adults. We selected asking the user to repeat their utterance (AskRepeat) as a baseline, and compared it to a combination of reprompting and a terse help message (ReprompThenSay, Reprompt + TerseYouCanSay from Table I). We chose these strategies because they are straightforward to implement and, unlike MoveOn, do not require educated guesses of good next topics to address. Specifically, our research questions are:

- **RQ1)** Which recovery strategy performs better, AskRepeat or ReprompThenSay?
- **RQ2)** Do older people benefit more than younger people from a given recovery strategy?

Our study was conducted online during the height of the Covid-19 pandemic using a partially simulated rule-based SDS implemented using PyDial [11]. This makes the setting slightly more realistic, because there was no experimenter

[1] School of Computing Science, University of Glasgow, Glasgow, UK Shaul.Ashkenazi@glasgow.ac.uk
[2] School of Informatics, University of Edinburgh, Edinburgh, UK

present to assist participants with technical difficulties. While rule-based systems are very rarely used in practice these days, they are ideal for controlled experiments with specific dialogue strategies. We discuss what LLM-based systems can learn from our findings.

II. Related Work

Dealing with errors has been investigated by the conversational user interface community for the past several decades. Much of this previous work has focused on spoken dialogue systems (SDS) with the classic five modules, automatic speech recognition (ASR), natural language understanding (NLU), dialogue management (DM), natural language generation (NLG), and text-to-speech (TTS). ASR converts spoken input into a representation that can be further processed by NLU. NLU extracts all relevant information from the utterance and passes it onto DM for further processing. DM controls the structure and flow of the dialogue, tracks information that has already been provided, and takes care of the interaction with relevant backend systems. Based on user input and backend data, it specifies the system's response, which is then converted into language by NLG and into speech by TTS. Recovery from problems is a key responsibility of the DM component.

Research within the traditional SDS literature established that systems should not just note that a problem has occurred. Instead, they should proactively work towards a joint solution, for example by asking a question [2], [12] or adding information [13].

One of the seminal papers on recovery strategies was Bohus and Rudnicky's comparison of ten different strategies for handling non-understanding errors [2]. They focused on first time users unfamiliar with how SDS work and non-native speakers, whose speech may be difficult for ASR systems to decode. They collected empirical data on the performance of the strategies and inferred appropriate error resolution strategies from their data. While simple reprompting worked comparatively well in Bohus and Rudnicky's study, other researchers found it to be less useful. The younger participants in Kim et al.'s study [14] favoured the simple AskRepeat strategy, because it clearly signalled that there had been a communication problem.

Opfermann and Pitsch [15] studied recovery strategies for three groups of participants: undergraduates (CTL), older adults with no cognitive impairment (SEN) and older adults with mild cognitive impairment (CIM). They found that CTLs produced by far more one-word turns than SENs and CIMs, with CIMs showing a decline with each issued reprompt. They concluded that reprompts are efficient if used once, but not multiple times. In addition, they proposed that after the first reprompt, followed by a user reaction, it would be best to use a "you can say" move (TerseYouCanSay in Table I).

Based on past studies, where AskRepeat as a baseline strategy was used (e.g., [13], [14], [4]), we decided to use it as a baseline in our study, and compare it to RepromptThenSay, a combination of Reprompt and TerseYouCanSay.

III. Methods

Ethical approval was granted by Tel Aviv University, Israel, where the first author conducted this research.

A. SDS and Recovery Strategies

Since it was not feasible to create an end-to-end rule-based SDS, we used a Wizard of Oz design [16]. In such studies, participants are told that they are interacting with a fully functional system, even though part of it is simulated by a human "wizard".

In our case, only the DM was simulated by the wizard. ASR was performed using Google services, while NLU was implemented using the PyDial toolkit [11]. The wizard then chose the ID of the appropriate response, which was played to the participant. The implemented recovery strategies are summarized in Figure 1.

1) **AskRepeat** – respond with the message *"Can you please repeat that?"*
2) **RepromptThenSay** (2-stage strategy):
   a) Repeat the question at a slower pace.
   b) Following a consecutive error, respond with the message: *"Sorry, I didn't catch that. You can say A, B or C"*, explicitly stating all the available options.

Fig. 1. Recovery Strategies Tested in This Study.

Participants were shown a red circle to tell them when they could start speaking. They were then recorded until the system detected 10 seconds of silence. This recording was then sent to the ASR system.

B. Participants

We recruited younger adults aged 18–30 and older adults aged 65 or above through ads in Facebook groups. All participants were native English speakers who spoke a variety of dialects of English.

C. Procedure

We used the restaurant booking domain for this study. Participants were asked to book a table at a restaurant for six different scenarios. Each scenario consisted of a date, a time, a dietary restriction, the party size, and a phone number for further questions. Scenarios 1, 3 and 4 invoked the AskRepeat strategy, while scenarios 2, 5 and 6 invoked the RepromptThenSay strategy.

The experiment was conducted via Zoom video chat during the Covid-19 pandemic.

1) Briefing: The participant was briefed on Zoom, to see that they are responsive and in a quiet environment.
2) Pretest: The participant was asked some questions to see if their equipment was in working condition.
3) Table-Reservation Scenarios: Participants performed six table reservation tasks, each described by a dedicated card presented in a fixed order for all participants.
4) SASSI: The participant was asked to fill a subset of the SASSI questionnaire [17].

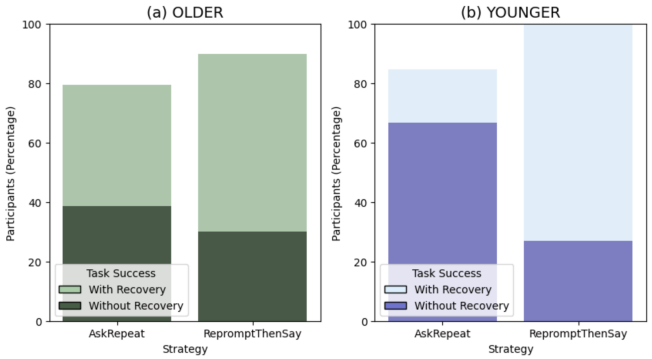

Fig. 2. Task Success

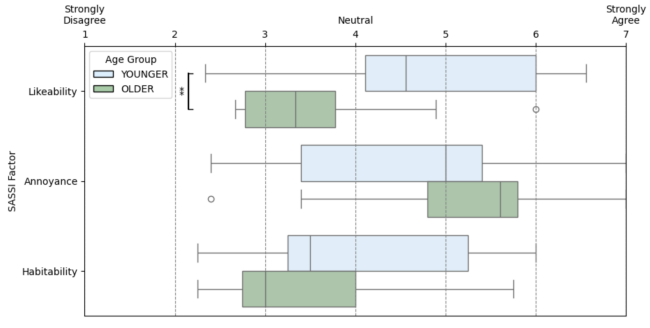

Fig. 3. SASSI Factor Scores by Age Group

5) Debriefing: The participant was asked about their experience and debriefed about the manipulation.

D. Evaluation Metrics

We used two metrics inspired by Bohus and Rudnicky [2]: 1) Task Success: a binary variable, which corresponded to whether a user had successfully booked a table within the constraints of the scenario card; 2) User Satisfaction: assessed using three factors of the SASSI questionnaire [17]: a) Likability: assesses whether the user finds the system useful, friendly, and pleasant; b) Annoyance: reflects negative perceptions, such as repetitiveness, boredom, irritation, and frustration; and 3) Habitability: reflects to what extent the user knows what to do and what the system is doing.

IV. Results

A. Demographics

We recruited a balanced sample of 26 people, 13 older and 13 younger. Older participants' mean age was 71 (range: 66–81 years) and younger participants' mean age was 26 (range: 22–29 years). Most participants were female. There were 3 men (23%) in the older age group, and 4 men (31%) in the younger age group. Six additional participants volunteered, but technical issues prevented their participation.

B. Task Success

Overall, 138 out of 156 tasks (156 = 2 age groups * 13 participants * 6 scenarios) were completed correctly, which corresponds to 88% of all tasks. Both older and younger participants successfully completed most tasks (Older: 66 out of 78, 85%; Younger: 72 out of 78, 92%). This difference in success rates is not significant (Wilcoxon signed rank test, paired, V = 42.5, $p = 0.1414$).

All 18 instances of failure involved an unsuccessful recovery attempt, 70 of the 138 successful completions involved a successful recovery, and 68 tasks were completed successfully without the need for recovery. This is reflected in Figure 2. Of the 18 unsuccessful recoveries, 14 (78%) used the AskRepeat strategy, while 4 (22%) used ReprompThenSay. All 4 participants who failed with ReprompThenSay were older. 6 younger and 8 older participants failed their task using AskRepeat.

A logistic regression shows that the effects of recovery ($p < 0.0001$) and strategy ($p = 0.01265$) are significant. There is no effect of scenario ($p = 0.12104$).[1] Thus, we conclude that ReprompThenSay outperforms AskRepeat. We did not have enough data to determine whether either of these strategies is particularly successful at supporting older participants.

C. Non-Understanding Errors: Sources

We labeled all interaction problems by source, with three categories from [2], and a fourth from our observations. All participants stayed within the restaurant reservation domain.

1) Out-of-Grammar: The participant used a word the SDS didn't recognize, e.g., *in the morning* instead of *AM*.
2) ASR: The participant's voice wasn't recognized correctly, generating the wrong text, e.g., *full* and not *four*.
3) End-pointer: A problem with the recording of the participant's voice, e.g., short or empty utterance.
4) Incorrect Data: Mismatch with card parameters.

Older users encountered more than twice as many issues as younger users, although the distribution of the problem types was quite similar. In both cases, most of the problems stemmed from out of grammar utterances. 8 older participants (61%) did not use the phrasing of X AM/PM. Instead, they used the phrases "in the morning/afternoon" or "o'clock".

The large number of ASR errors may be due to the range of English dialects represented in the sample, in addition to specific pronunciation issues. E.g., some participants struggled with words like "halal" for dietary restrictions, while others used idiosyncratic pronunciations of numbers.

Finally, technical issues accounted for a few problems. In addition to network latency, sometimes, participants spoke too early or too late, which meant that their voices were not recorded at all, which led to overall frustration.

D. User Satisfaction

Figure 3 shows the overall SASSI scores of older and younger participants. Younger and older people were equally

[1]To compute p-values, we used stepwise model comparison with analysis of deviance and the likelihood ratio test. Thus, the effect of recovery is determined by comparing a model consisting just of the intercept with a model consisting of intercept plus recovery, and so on.

annoyed at the experience of interacting with the table reservation system (older: M=5.17; younger: M=4.52, $t(24) = 1.20, p = .24, d = 0.47$), and its habitability was average (older: M=3.46; younger: N=4.13, $t(24) = 1.50, p = .14, d = 0.60$). Younger users (M=4.83) rated the system overall as more likeable than older users (M=3.58, $t(24) = 2.90, p = .0087, d = 1.12$)

E. Different Past User Experiences

Younger participants were more likely than older ones to give several of the reservation parameters in the first message (8 vs. 4). This may reflect greater familiarity with using SDS and assuming it could handle the information.

V. Discussion and Conclusion

With regard to RQ1, we found that RepromptThenSay outperformed AskRepeat as a recovery strategy. Due to the overall high task success on the table reservation task, and the relatively small effect of the recovery strategy, we were unable to conclude which strategy would be more suitable to older participants (RQ2). Instead, what proved critical was having any recovery strategy at all: all failed interactions involved a failed recovery, while half of the successful ones succeeded due to a successful recovery. Rather than inducing problems, we observed how our strategies handled naturally occurring issues.

Our study had several limitations. The online setup introduced unnatural turn-taking and latency. Older participants were likely more tech-literate than average, potentially inflating success rates. A less tech-savvy group might show lower success and greater need for recovery. We would also have benefited from a larger sample size, given the small effect of strategy. Finally, we did not inform users about the recovery strategies, so no direct feedback was collected.

Perhaps the main limitation is our use of a rule-based SDS rather than an LLM-based chatbot or a more sophisticated statistical SDS, due to the study's timing. Still, we believe our findings apply to LLMs: effective recovery strategies remain crucial. Although LLMs more or less eliminate the issue of out of grammar errors, problems like incorrect input and speech recognition errors persist.

While LLMs use a range of recovery strategies, prompting them with those proven effective may improve performance. This can be coupled with instructions that alert LLMs to specific user challenges. For instance, users with cognitive decline struggle with self-correction [18]. In order to create truly inclusive voice assistants, we suggest that future work should focus more on supporting specific challenges such as cognitive decline, specific preferences for social and supportive user interfaces [19], or a tendency to get easily frustrated with conversational user interfaces.

References

[1] S. McROY, "Preface: Detecting, repairing and preventing human—machine miscommunication," *International Journal of Human-Computer Studies*, vol. 48, no. 5, pp. 547–552, 1998.

[2] D. Bohus and A. I. Rudnicky, "Sorry, i didn't catch that! an investigation of non-understanding errors and recovery strategies," *Recent trends in discourse and dialogue*, pp. 123–154, 2008.

[3] A. Mahmood, J. W. Fung, I. Won, and C.-M. Huang, "Owning mistakes sincerely: Strategies for mitigating ai errors," in *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, 2022, pp. 1–11.

[4] A.-M. Meck, C. Draxler, and T. Vogt, "Failing with grace: Exploring the role of repair costs in conversational breakdowns with in-car voice assistants," *International Journal of Human–Computer Interaction*, vol. 40, no. 22, pp. 7574–7592, 2024.

[5] D. Benner, E. Elshan, S. Schöbel, and A. Janson, "What do you mean? a review on recovery strategies to overcome conversational breakdowns of conversational agents." in *ICIS*, 2021.

[6] R. Brewer, C. Pierce, P. Upadhyay, and L. Park, "An empirical study of older adult's voice assistant use for health information seeking," *ACM Transactions on Interactive Intelligent Systems (TiiS)*, vol. 12, no. 2, pp. 1–32, 2022.

[7] Z. Yang, X. Xu, B. Yao, E. Rogers, S. Zhang, S. Intille, N. Shara, G. G. Gao, and D. Wang, "Talk2care: An llm-based voice assistant for communication between healthcare providers and older adults," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 8, no. 2, pp. 1–35, 2024.

[8] P. Upadhyay, S. Heung, S. Azenkot, and R. N. Brewer, "Studying exploration & long-term use of voice assistants by older adults," in *Proceedings of the 2023 CHI conference on human factors in computing systems*, 2023, pp. 1–11.

[9] M. Geng, X. Xie, Z. Ye, T. Wang, G. Li, S. Hu, X. Liu, and H. Meng, "Speaker adaptation using spectro-temporal deep features for dysarthric and elderly speech recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 2597–2611, 2022.

[10] M. Wolters, K. Georgila, J. D. Moore, and S. E. MacPherson, "Being old doesn't mean acting old: How older users interact with spoken dialog systems," *ACM Transactions on Accessible Computing (TACCESS)*, vol. 2, no. 1, pp. 1–39, 2009.

[11] S. Ultes, L. M. R. Barahona, P.-H. Su, D. Vandyke, D. Kim, I. Casanueva, P. Budzianowski, N. Mrkšić, T.-H. Wen, M. Gasic *et al.*, "Pydial: A multi-domain statistical dialogue system toolkit," in *Proceedings of ACL 2017, System Demonstrations*, 2017, pp. 73–78.

[12] G. Skantze, "Exploring human error handling strategies: Implications for spoken dialogue systems," in *ISCA Tutorial and Research Workshop on Error Handling in Spoken Dialogue Systems*. Centre for Speech Technology KTH Stockholm, Sweden, 2003, pp. 71–76.

[13] M. Henderson, C. Matheson, and J. Oberlander, "Recovering from non-understanding errors in a conversational dialogue system," in *Proceedings of SemDial 2012 (SeineDial): The 16th Workshop on the Semantics and Pragmatics of Dialogue*, vol. 128, 2012.

[14] J. Kim, M. Jeong, and S. C. Lee, ""Why did this voice agent not understand me?" error recovery strategy for in-vehicle voice user interface," in *Proceedings of the 11th International Conference on Automotive User Interfaces and Interactive Vehicular Applications: Adjunct Proceedings*, 2019, pp. 146–150.

[15] C. Opfermann and K. Pitsch, "Reprompts as error handling strategy in human-agent-dialog? user responses to a system's display of non-understanding," in *2017 26th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*. IEEE, 2017, pp. 310–316.

[16] N. Dahlbäck, A. Jönsson, and L. Ahrenberg, "Wizard of oz studies: why and how," in *Proceedings of the 1st international conference on Intelligent user interfaces*, 1993, pp. 193–200.

[17] K. S. Hone and R. Graham, "Towards a tool for the subjective assessment of speech system interfaces (sassi)," *Natural Language Engineering*, vol. 6, no. 3-4, pp. 287–303, 2000.

[18] M. Kobayashi, A. Kosugi, H. Takagi, M. Nemoto, K. Nemoto, T. Arai, and Y. Yamada, "Effects of age-related cognitive decline on elderly user interactions with voice-based dialogue systems," in *Human-Computer Interaction–INTERACT 2019: 17th IFIP TC 13 International Conference, Paphos, Cyprus, September 2–6, 2019, Proceedings, Part IV 17*. Springer, 2019, pp. 53–74.

[19] S. Kopp, M. Brandt, H. Buschmeier, K. Cyra, F. Freigang, N. Krämer, F. Kummert, C. Opfermann, K. Pitsch, L. Schillingmann *et al.*, "Conversational assistants for elderly users–the importance of socially cooperative dialogue," in *Proceedings of the AAMAS Workshop on Intelligent Conversation Agents in Home and Geriatric Care Applications co-located with the Federated AI Meeting*, vol. 2338, 2018.