Inclusive needs assessments should be the foundation for motivating technical CUI research

Andrea Cuadra
Stanford University

Jackie (Junrui) Yang
Stanford University

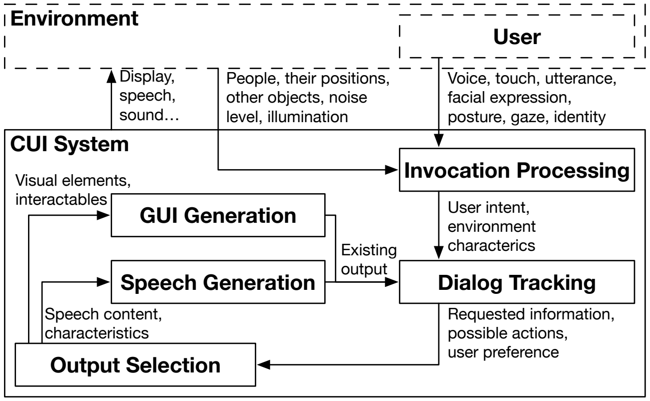
Environment
User
Display, speech, sound…
People, their positions, other objects, noise level, illumination
Voice, touch, utterance, facial expression, posture, gaze, identity
CUI System
Invocation Processing
Visual elements, interactables
GUI Generation
User intent, environment characterics
Existing output
Speech Generation
Dialog Tracking
Speech content, characteristics
Requested information, possible actions, user preference
Output Selection

Figure 1: System Diagram. All information flows depicted could be made more inclusive. Note, this list of information flows not comprehensive.

Abstract

Including older adults in the design of future Conversational User Interfaces (CUIs) is not only the *right* thing to do, but the *smart* thing to do when designing across modalities and mobilities. If you strive to meet the needs and preferences of as diverse an audience as older adults, then you can end up with adaptable and delightful conversation-first interfaces. In this position paper, we present cutting-edge research that shows how the findings from an in-depth video analysis of older adults' interactions with CUIs can deeply enhance technical research that pushes the interaction boundaries of CUIs.

Author Keywords

Conversational User Interfaces; Voice-First Ambient Interfaces; smart speakers; multimodal devices; voice assistants

Introduction

It is well-accepted in the HCI community that CUIs have great potential for serving older adults [11, 7, 16, 8, 12, 1], but that CUIs' current limitations create barriers to realizing this potential [13, 4]. These barriers affect other users as well [2, 3]. To mitigate challenges that are often encountered in interactions with CUIs, it is crucial to design them considering the needs of a wide range of users.

In this position paper, we argue that designing CUIs to meet the needs and preferences of older adults can help prioritize work based on the potential for positive impact, and generate novel and meaningful research ideas for future research. The authors of this position paper are both HCI researchers, one whose expertise is on building systems to support a wide range of interactions, and the other whose expertise is on inclusive design.

Interaction Needs and Proposed Mechanisms

In this section, we describe a set of older adult interaction needs identified in prior work [4], and describe how existing research about sensing and interactions (e.g., [14, 15]) may be used to address to needs. Prior work employing video analysis to examine the use of a smart speaker by older adults identified a series of interaction needs older adults have in these interactions [4]. We now describe specific needs and proposed mechanisms to address them.

Participants in [4] attempted to talk to the smart speaker, but, for many reasons (e.g., omitting saying the wake word at the beginning of an utterance), their interaction attempts were overlooked. This portrays an interaction need for alternative waking mechanisms. We list some:

1. By using Soundr [14] to detect body language and position in space, such as head orientation and proximity to the system, that consistently signals an interaction attempt.

2. Some participants used alternate pronunciations to Alexa, such as Alessa, or Alexia. Foundation models could be used to detect interaction attempts in a way that more closely mimics human interactions, such as by detecting close enough words to the “wake” word.

3. Several participants interacted in groups of more than one. The system could use the mechanisms described in [14, 10], speaker recognition, image classification, and information from other connected devices to detect how many users are interacting with it and tailor the interactions accordingly.

4. Using alternative, more intuitive, or more familiar, design paradigms, such as using a landline phone instead of a wake word could also help disambiguate these interactions. This could be set up in many ways, such as by re-purposing old phones by setting them up with build-in voice assistants or by using a smart speaker’s existing sensors (e.g., camera, microphone) in addition to computer vision or sound recognition to detect that the phone has been picked up.

Together, this list shows many potential solutions that would address many of the conversational breakdowns observed by Cuadra et al. [4], which happened at the very beginning of an interaction attempt, where the system failed to detect a trigger.

Moreover, there are many other ways to make the response of the system more inclusive. For example, systems can mirror and understand users, such as setting the system’s speed and intonation based on the user’s, to adapt to a user’s needs and abilities. While some participants spoke slowly, loudly, or with a lot of pauses, others did the opposite. In addition, some participants during some interactions said many words to the voice assistant during a single interaction, sometimes speaking for longer than the voice assistant could process. Using this information to inform the system’s reactions and expectations could help improve the quality and accessibility of the interactions. Prior research also indicates that by adjusting the system’s intonation, the system can appear more trustworthy [9]. This finding creates a novel idea for future research using tone analysis as input and attributes of speech synthesis as output. Apart from speech characteristics, future systems should also adapt content to the users' varying capabilities and preferences. On the graphics user interface side, SUPPLE [5] has investigated adjusting the actual control (drop-down box vs option box group), font size, and control size to use based on a user's ability. We think future research can address this on a cross-device conversational user interface with similar generative UI techniques. The needs could again, be recognized using visual and audio information gathered from the system's sensors.

Finally, participants in [4] were confused when they pointed or reacted to something on the smart speaker's screen and the smart speaker did not take into account what it was displaying as part of the dialog state. Multimodal interactions need to be consistent with each other, and respond in intuitive ways for the user. To this end, we could employ the technology described in [15, 6] to meet this interaction need, and evaluate whether this interaction need could be supported accordingly.

Proposed system

To make the process of addressing these interactions, we created a system diagram in Figure 1 that illustrates some of the main information flows that need to be considered when designing CUIs inclusively. In the diagram, the labels for the arrows describe information flows that will dynamically vary by a specific user, specific situation or environment, and type of interaction that the user is intending to have; and the blocks represent the different components of a proposed CUI system can potentially have.

The system starts by detecting an invocation in a multimodal adaptive way. It processes the user's interactions with the system (e.g., voice, utterance, and text), and captures the user's auxiliary information (e.g., facial expression, posture, etc.). It also monitors the user's environment (e.g., ambient sound, light, etc.) and the social context (e.g., the number of people in the room). With this information, the system can engage when is useful and appropriate. For example, the system may allow a long pause when the user is talking slowly, and it would have a faster response when it detects a fast-talking user.

The system then abstracts the user's intent and environment characteristics and uses a dialog tracking system to handle the user's request within the context of the existing output and the current environment. Once the system retrieves the information, the system selects the appropriate output based on the user's preferences and context. It can make sure that it answers the user's question in a way that is understandable and appropriate for the user. Finally, the CUI system would then generate the output and present it through display, speech, and sound. With this approach, the system would select the modality that is better suited for the user's situation: making sure not to disrupt people around the user and presenting the information in a format that is easily accessible to the user.

Conclusion

In this position paper, we analyzed how findings from a needs assessment of older adults' interactions with multimodal smart speakers can be used to inform, motivate, and apply other technical research towards making CUIs usable by a wider range of users. In carrying out this analysis, we found that considering the needs of older adults helps steer future research in novel directions with a high potential for positive impact.

REFERENCES

[1] Sarah Abdi, Luc de Witte, and Mark Hawley. 2020. Emerging Technologies With Potential Care and Support Applications for Older People: Review of Gray Literature. *JMIR Aging* 3, 2 (aug 2020), e17286. DOI: http://dx.doi.org/10.2196/17286

[2] Erin Beneteau, Olivia K. Richards, Mingrui Zhang, Julie A. Kientz, Jason Yip, and Alexis Hiniker. 2019. Communication Breakdowns Between Families and Alexa. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM. DOI: http://dx.doi.org/10.1145/3290605.3300473

[3] Leigh Clark, Nadia Pantidi, Orla Cooney, Philip Doyle, Diego Garaialde, Justin Edwards, Brendan Spillane, Emer Gilmartin, Christine Murad, Cosmin Munteanu, Vincent Wade, and Benjamin R. Cowan. 2019. What Makes a Good Conversation?: Challenges in Designing Truly Conversational Agents. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM. DOI: http://dx.doi.org/10.1145/3290605.3300705

[4] Andrea Cuadra, Hyein Baek, Deborah Estrin, Malte Jung, and Nicola Dell. 2022. On Inclusion: Video Analysis of Older Adult Interactions with a Multi-Modal Voice Assistant in a Public Setting. In *International Conference on Information & Communication Technologies and Development (ICTD)*.

[5] Krzysztof Z. Gajos, Jacob O. Wobbrock, and Daniel S. Weld. 2008. Improving the performance of motor-impaired users with automatically-generated, ability-based interfaces. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM. DOI: http://dx.doi.org/10.1145/1357054.1357250

[6] Michael Johnston, John Chen, Patrick Ehlen, Hyuckchul Jung, Jay Lieske, Aarthi Reddy, Ethan Selfridge, Svetlana Stoyanchev, Brant Vasilieff, and Jay Wilpon. 2014. MVA: The Multimodal Virtual Assistant. In *Proceedings of the 15th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL)*. Association for Computational Linguistics. DOI: http://dx.doi.org/10.3115/v1/w14-4335

[7] Sunyoung Kim and Abhishek Choudhury. 2021. Exploring older adults' perception and use of smart speaker-based voice assistants: A longitudinal study. *Computers in Human Behavior* 124 (nov 2021), 106914. DOI: http://dx.doi.org/10.1016/j.chb.2021.106914

[8] Aqueasha Martin-Hammond, Sravani Vemireddy, and Kartik Rao. 2019. Exploring Older Adults' Beliefs About the Use of Intelligent Assistants for Consumer Health Information Management: A Participatory Design Study. *JMIR Aging* 2, 2 (dec 2019), e15381. DOI:http://dx.doi.org/10.2196/15381

[9] Katherine Metcalf, Barry-John Theobald, Garrett Weinberg, Robert Lee, Ing-Marie Jonsson, Russ Webb, and Nicholas Apostoloff. 2019. Mirroring to Build Trust in Digital Assistants. In *Interspeech 2019*. ISCA. DOI:http://dx.doi.org/10.21437/interspeech.2019-1829

[10] O. Miksik, I. Munasinghe, J. Asensio-Cubero, S. Reddy Bethi, S-T. Huang, S. Zylfo, X. Liu, T. Nica, A. Mitrocsak, S. Mezza, R. Beard, R. Shi, R. Ng, P. Mediano, Z. Fountas, S-H. Lee, J. Medvesek, H. Zhuang, Y. Rogers, and P. Swietojanski. 2020. Building Proactive Voice Assistants: When and How (not) to Interact. (2020). DOI: http://dx.doi.org/10.48550/ARXIV.2005.01322

[11] Alisha Pradhan, Amanda Lazar, and Leah Findlater. 2020. Use of Intelligent Voice Assistants by Older Adults with Low Technology Use. *ACM Transactions on Computer-Human Interaction* 27, 4 (aug 2020), 1–27. DOI:http://dx.doi.org/10.1145/3373759

[12] Sergio Sayago, Barbara Barbosa Neves, and Benjamin R Cowan. 2019. Voice assistants and older people: some open issues. In *Proceedings of the 1st International Conference on Conversational User Interfaces*. ACM. DOI: http://dx.doi.org/10.1145/3342775.3342803

[13] Milka Trajkova and Aqueasha Martin-Hammond. 2020. "Alexa is a Toy": Exploring Older Adults' Reasons for Using, Limiting, and Abandoning Echo. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. ACM. DOI: http://dx.doi.org/10.1145/3313831.3376760

[14] Jackie (Junrui) Yang, Gaurab Banerjee, Vishesh Gupta, Monica S. Lam, and James A. Landay. 2020a. Soundr: Head Position and Orientation Prediction Using a Microphone Array. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. ACM. DOI: http://dx.doi.org/10.1145/3313831.3376427

[15] Jackie (Junrui) Yang, Monica S. Lam, and James A. Landay. 2020b. DoThisHere: Multimodal Interaction to Improve Cross-Application Tasks on Mobile Devices. In *Proceedings of the 33rd Annual ACM Symposium on User Interface Software and Technology*. ACM. DOI:http://dx.doi.org/10.1145/3379337.3415841

[16] Tamara Zubatiy, Kayci L Vickers, Niharika Mathur, and Elizabeth D Mynatt. 2021. Empowering Dyads of Older Adults With Mild Cognitive Impairment And Their Care Partners Using Conversational Agents. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. ACM. DOI: http://dx.doi.org/10.1145/3411764.3445124