Situated Conversational Agents for Task Guidance: A Preliminary User Study

Alexandra Bremers*†
Cornell Tech
New York, USA
awb227@cornell.edu

Manaswi Saha*
Accenture Labs
San Francisco, USA
manaswi.saha@accenture.com

Adolfo G. Ramirez-Aristizabal
Accenture Labs
San Francisco, USA
adolfo.ramirez@accenture.com

ABSTRACT

Multimodal large language models have enabled a new generation of Conversational Agents (CA), leveraging language structure in human discourse to encode-decode multimedia formats (*e.g.,* video-to-audio). These next-generation CAs can be useful in task guidance scenarios, where the user's attention space is limited and verbal instructions can be overwhelming. In this paper, we explore the role of non-verbal conversational cues in identifying and recovering from errors while performing various assembly tasks. Findings from an exploratory Wizard-of-Oz study (*N*=8) indicate individual differences and preferences for auditory guidance. Combining these initial findings with our early exploration of the task monitoring system, we discuss implications for the emerging area of situated multimodal CAs for physical task guidance, where conversational interactions are based on inputting visual task actions and generating auditory feedback.

KEYWORDS

multimodal, task guidance, physical assembly, audio augmented reality, conversational agents

ACM Reference Format:
Alexandra Bremers, Manaswi Saha, and Adolfo G. Ramirez-Aristizabal. 2024. Situated Conversational Agents for Task Guidance: A Preliminary User Study. In *ACM Conversational User Interfaces 2024 (CUI '24), July 8–10, 2024, Luxembourg, Luxembourg*. ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3640794.3665575

1 INTRODUCTION

The advent of Large Language Models (LLMs) such as GPT-4o [1] and Astra[2] have ushered the development of a new generation of Conversational Agents (CAs), enabling multimodal interactions by processing images, video, sound, and text both as input and output [34]. Thus, allowing CAs to have new context-aware capabilities that can generalize beyond traditional text-based conversations [18]. In studies investigating audio-based task-guidance systems [22], researchers have explored context-aware systems that analyze semantic relationships in visual scenes and translate that back as auditory guidance to a user, which forms the basis of this work (Figure 1).

For industrial applications (*e.g.,* manufacturing, healthcare), user cognitive load and task performance are factors that directly guide design principles [4]. Therefore, task guidance seeks to balance the informational bandwidth of users, in which speech feedback is often found to be distracting when a user is focused on a coordinated physical task [33]. In this paper, we seek to advance the under-explored area of non-verbal auditory guidance by CAs. In the exploratory Wizard-of-Oz (WoZ) [8] study, participants assemble various home healthcare devices (*e.g.,* CPAP mask) using discrete and continuous auditory guidance. We also built a WoZ-based task monitoring system that captures the user performing the task as well as the table-top task (Figure 2a, Figure 5). Through semi-structured interviews and self-reported scores of mental effort during tasks using the guidance modes, our findings show initial trends in user preferences in auditory guidance feedback. Using these preliminary findings, we make a case for CA systems to include *video-audio situated CAs* for physical task guidance and present a preliminary discussion on designing these systems.

2 BACKGROUND

Below, we provide a background on situated multimodal conversational systems and their affordances as they relate to physical task guidance and present the vision for such systems.

2.1 Physical Task Guidance

Physical tasks (*e.g.,* assembly) across different industries (*e.g.,* manufacturing, healthcare) are complex in nature [24, 30], with multiple components in play, ranging from simpler self-assembly products (*e.g.,* home healthcare devices) to complex assembly line tasks (*e.g.,* 80-component lift [3]). They are often conducted in a high stakes, stressful environments (*e.g.,* factories, hospitals), where users (*e.g.,* workers, medical professionals) have high demands for visual attention and engagement as well as high task performance and execution needs [4, 38]. For these environments, mixed-initiative task-oriented systems [13, 14, 40] would work best. This task guidance context can be modeled as a multimodal conversation (Figure 1): the inputs are sensed from the environment (*e.g.,* user task actions, sound profile), and guidance is generated considering the needs of the physical context (*e.g.,* noisy), task context (*e.g.,* task complexity) and the user context (*e.g.,* worker expertise) providing a hands-free, eyes-free interaction experience [32]. With this vision in mind, we explore the relevance of situated multimodal CAs for physical task guidance, where the task action video is the basic unit of analysis, instead of speech as in traditional CAs.

*Both authors contributed equally to the paper.
†Work done while at Accenture Labs.

CUI '24, July 8–10, 2024, Luxembourg, Luxembourg

ACM ISBN 979-8-4007-0511-3/24/07.
https://doi.org/10.1145/3640794.3665575

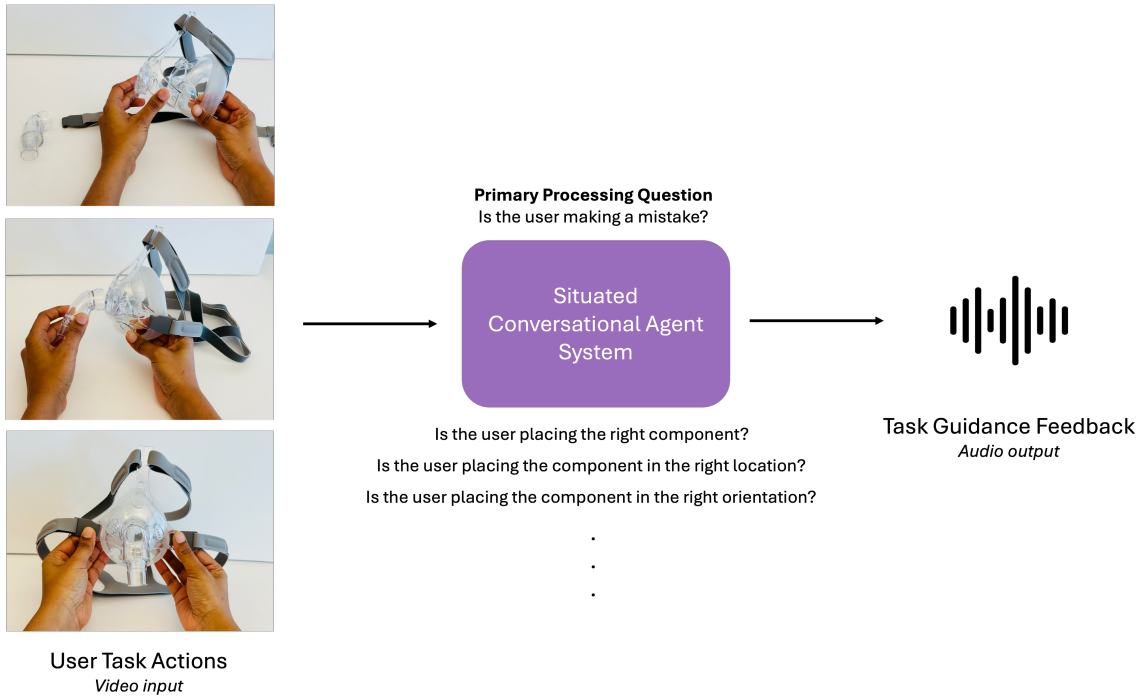


Figure 1: Illustration of task guidance as a multimodal conversation: user action (in a visual scene) is fed into the situated CA system, with task progress and error detection questions as the queries and task guidance via auditory feedback as the generated output.

2.2 Situated Multimodal Conversations and Agents

Human conversations are complex and multimodal, consisting of verbal and non-verbal cues [11, 25]. Beyond speech, multimodal conversations rely on interaction modalities such as gestures (*e.g.,* sign language), facial expressions, emotions, and body postures. Multimodal interactions enhance performance by providing humans alternative modalities, when some are busy (*e.g.,* occupied visual field) [23]. To support these interactions, multimodal CAs (*e.g.,* coaches [10]) take various input modalities (*e.g.,* images, videos), but have traditionally generated text [26]. Closest to our context are situated CAs [7, 17, 21, 27, 37], which are a type of multimodal task-oriented dialog system [39] that encompass a situated multimodal user-task context, handling multiple input-output modalities where context is defined by the affordances of the task and embodiment of the CA. In this work, we focus on *video-audio situated CAs*, where user task action video is used as an input to generate non-speech auditory feedback, which we refer to as non-verbal cues in the paper. We consider single input and output modality for this work. Handling multiple modalities at the same time is out of scope.

For physical task guidance, systems have traditionally focused on text-oriented Q&A (*e.g.,* LLM-based cognitive assistants [15]) or visual guidance (*e.g.,* traditional AR [19, 29, 35, 36]). However, due to the visual attention constraints of physical tasks, auditory feedback is cognitively less taxing [28, 31]. Compared to verbal (speech) auditory feedback (*e.g.,* spoken dialog systems [40]), non-verbal cues have lower cognitive load, especially in task guidance scenarios [6]. A growing body of research has focused on audio Augmented Reality (AR) solutions that utilize non-verbal cues and sonification techniques to augment the physical environment [9, 12]. In this work, we are investigating the role of non-verbal cues in a video-based task-oriented dialog system [34], an underexplored area for augmented environments [20]. We take inspiration from earcons [5] and common intuitive non-verbal cues for guidance sounds: discrete beeps for status (*e.g.,* microwave) and continuous tones for progress (*e.g.,* parking sensor). Furthermore, we specifically focus on error recovery guidance, as opposed to providing step-by-step instructions.

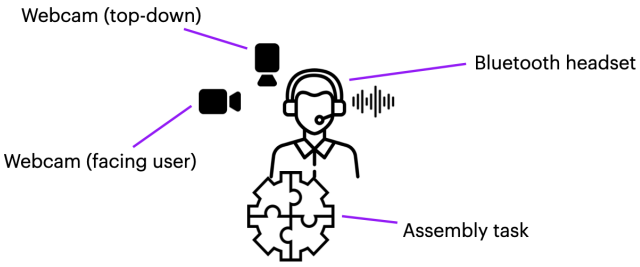
3 USER STUDY

3.1 Study Design

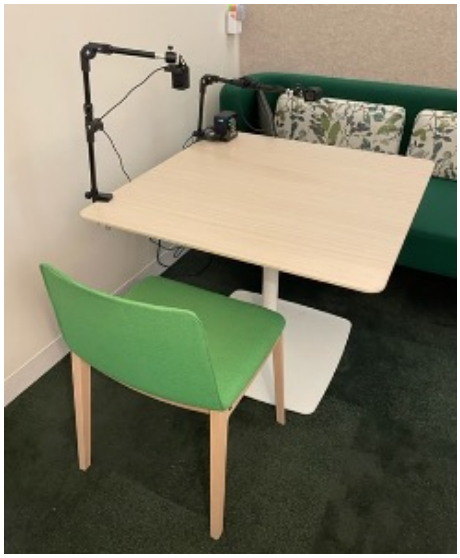
The aim of the exploratory Wizard-of-Oz study is to identify the design opportunities of non-verbal auditory task guidance. The protocol contains three phases: 1) familiarization with a physical assembly task and the guidance system, 2) performing three assembly tasks with three guidance modes, and 3) semi-structured interviews to elicit feedback on their experience. The experimental design consisted of nine (3X3) conditions. The audio guidance modes are continuous tones, discrete earcons, and a control condition of no guidance. Continuous mode played a tone throughout the task, but increased the sound volume during an error and lowered it once corrected. Discrete tones used specific earcons to indicate each task progress state: error, on track, step complete, task complete.

3.1.1 Tasks. The study protocol had an introductory task and three study tasks (Figure 3). The introductory task consisted of a Lego assembly task where participants had to add the two missing pieces to create the feet of a Lego man. During this phase, they try out the discrete and continuous guidance modes by intentionally making mistakes and listening to the generated audio cues. The subsequent three condition tasks included assembling a manual breast pump, a CPAP mask, and a heart-shaped Lego structure (Figure 3). The Lego condition was chosen because it is a general and application agnostic task while the other two conditions are analogues to home healthcare assembly. For each of these three tasks, participants would hear one sound condition: control (no guidance), continuous guidance, or discrete guidance. The task order was randomized (Table 1).

3.1.2 Instruction Videos. Before each task, the participant watched a video showing a tabletop view of someone performing the assembly task. The video was played once, without audio. The videos were created by the authors recording their own hands while performing assembly.



(a) Wizarding system. A participant receives auditory cues while completing a tabletop-based physical assembly task, monitored by the cameras. (Icons via Flaticon.)



(b) Room Setup

Figure 2: Wizarding System Setup

Table 1: Order of Task and Condition per Participant Condition Types: Disc = Discrete Guidance, Cont = Continuous Guidance, None = No Guidance.

P#	T1_CondType	T2_CondType	T3_CondType
P1	Lego_None	BPump_Disc	CPAP_Cont
P2	Lego_None	BPump_Disc	CPAP_Cont
P3	CPAP_Cont	BPump_Disc	Lego_None
P4	BPump_Cont	CPAP_Disc	Lego_None
P5	CPAP_Cont	BPump_Disc	Lego_None
P6	BPump_Cont	CPAP_Disc	Lego_None
P7	CPAP_Disc	BPump_Cont	Lego_None
P8	BPump_Disc	CPAP_Cont	Lego_None

3.1.3 Guidance Sounds. We designed the sound cues based on guidelines by Blattner et al. [5], which state that earcons should be simple, ideally taken from the same octave. We focused on designing the continuous sounds as ambient, slightly alerting presence; the discrete sounds as affirming, encouraging, yet alerting in case of a mistake. The sounds were designed in GarageBand with the musical typing virtual MIDI keyboard function. The resulting sounds are provided as supplementary materials.

3.1.4 Wizard-of-Oz Setup. The platform consisted of two cameras (Figure 2a): one observed the participant and one was pointed towards the work surface. A Bluetooth headset was connected to the computer that played sounds for guidance, triggered by the wizard. Our system was built using a Raspberry Pi, a laptop, a tablet, and an external monitor. The WoZ task guidance system was set up in a 10' x 6'6" room. The wizard was present in the same room, but the table obstructed the participant's view of what the wizard was controlling (Figure 2b).

3.2 Participants

Eight participants (4F, 3M, 1 unanswered), with an average age of 37.4 years (range 27 to 57), were recruited from Accenture. All participants provided informed consent and received a gift voucher for participation. At the start of the study, participants were asked about their experience with Legos, Breast Pumps, and CPAP masks. Participants generally had the highest familiarity with Lego and the lowest familiarity with the CPAP mask (Figure 4). The protocol was reviewed and approved by Accenture Legal.

3.3 Procedure

Participants wore latex gloves, were handed the Bluetooth headset, and were asked about the comfort level of the volume. The participants completed demographics and experience questionnaires, ranking their familiarity with the task and their commonly used instruction modalities. Participants then performed the introductory task while experiencing the two guidance modes, followed by the three study tasks. After each study task, the participants completed the NASA TLX and shared their immediate reactions to the task. Finally, participants participated in a semi-structured debrief interview, asking about their experience on the tasks and the guidance modes.

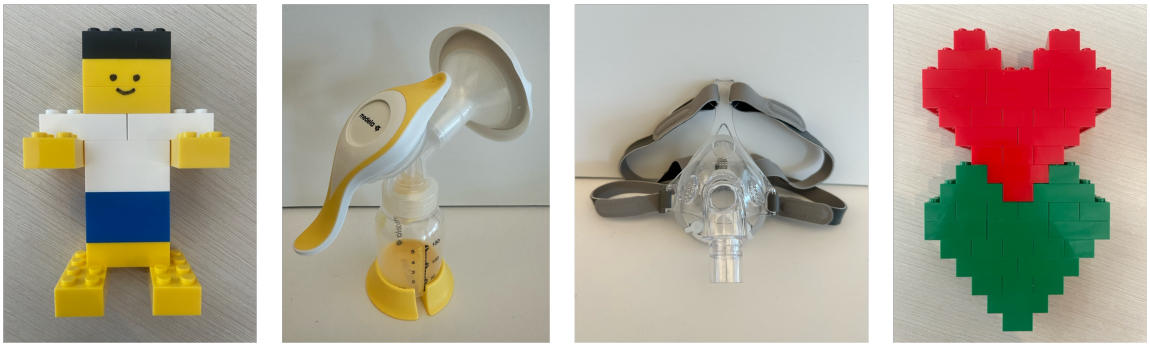


Figure 3: Introductory task: Lego Man (leftmost). Three study tasks (L-R): Manual Breast Pump, CPAP Mask, Lego Heart.

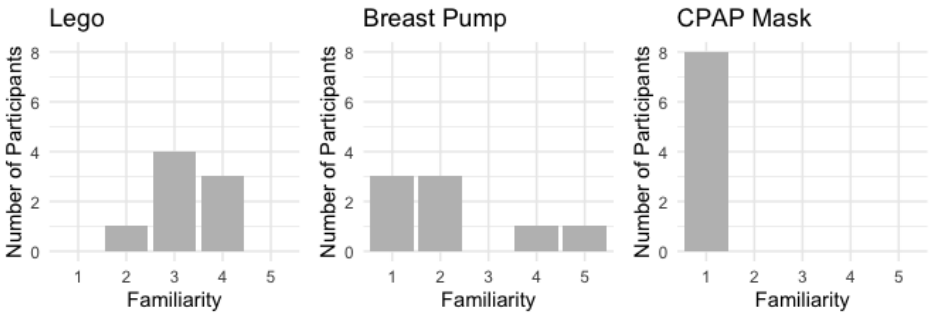


Figure 4: Self-reported familiarity with each task, from 1="I don't know what it is" to 5="I'm an expert on this".

4 FINDINGS

We discuss insights from the semi-structured interviews and self-reported NASA TLX scores about participants' experiences during tasks and preferences of different guidance modes.

4.1 Task Experience

Participants stated that the difficulty increased with tasks with more homogeneous steps: *"I think the Lego heart was the most difficult because all the pieces were the exact same, and they had a lot of possible orientations. [...] The breast pump and the CPAP mask had only 4 to 5 distinct parts at most, all of which looked very different from each other."* (P1). Furthermore, the unfamiliar equipment caused nervousness: *"With respect to the mask I was very much afraid that I'm gonna break some piece."* (P3). All participants were least familiar with assembling a CPAP mask, compared to the other tasks (Figure 4). Finally, some participants decided not to follow the step order from the instruction video: *"During the tasks, I had to keep in mind: what was the order of [steps] in the video, and how is that different from how I think? The order of the assembly should work for me, so I made a conscious decision of not following the exact order [of the instruction video]."* (P6).

4.2 Guidance Mode Experience

4.2.1 Audio as a Feedback Modality. With audio feedback, participants felt more confident and encouraged: *"[For the discrete guidance] I think just because of the reinforcement as I went along, it gave me confidence that I was moving towards the correct direction."* (P5). P4 added that their stress level reduced: *"Having some guidance makes me feel less stressful, a bit more confident compared to no guidance situations."* Audio feedback also helped them focus: *"It changed the rhythm of my assembly. [...] If I, for example, became disengaged, I could easily ground myself back in the assembly task. In the ones where audio guidance was sparse, there was less constant grounding."* (P6). When comparing the two guidance modes, participants found the tonal characteristics of the discrete tones calmer than the continuous one: *" I didn't like the continuous one so much just because it was a little loud. A little overwhelming. Discrete was much calmer, I felt like it was encouraging me, but the continuous was yelling at me."* (P5).

4.2.2 Interpreting the guidance modes. Participants' unweighted NASA TLX scores (Table 2) indicate that tasks with discrete guidance were perceived as less demanding, compared to continuous and no guidance, across all dimensions. Forgoing statistical significance testing due to the small sample size, these initial trends suggest that discrete guidance may provide a more user-friendly and efficient task experience. However, participants reported mixed preferences and opinions on interpretability. For example, P1 preferred discrete over continuous because of the perceived direct relation with an action: *"I definitely prefer the discrete guidance, the little "pings", because it responded directly to an action that I would do. With the sound increasing or decreasing, I couldn't really tell what decision or what choice or what task it was responding to, so it was less like instant feedback and less clear."* In contrast, some participants found the sound changes in the continuous mode easier to detect: *"I*

Table 2: NASA TLX Dimensions (range 0-100)

Guidance	TLX Dimension	Mean	SD
Continuous	Effort	40.62	25.42
Continuous	Frustration Level	28.75	24.89
Continuous	Mental Demand	39.38	24.99
Continuous	Performance	14.38	21.95
Continuous	Physical Demand	13.75	16.42
Continuous	Temporal Demand	25.00	24.35
Discrete	Effort	17.50	13.36
Discrete	Frustration Level	10.62	7.29
Discrete	Mental Demand	19.38	14.74
Discrete	Performance	5.62	5.63
Discrete	Physical Demand	10.62	9.43
Discrete	Temporal Demand	11.88	13.08
None	Effort	42.50	19.27
None	Frustration Level	38.75	31.02
None	Mental Demand	45.62	22.11
None	Performance	17.50	16.90
None	Physical Demand	27.50	23.30
None	Temporal Demand	35.62	29.33

liked the continuous sound because it was easy for me to understand if it's changing in modulation." (P3).

However, even with audible sound changes, it was hard to remember the meaning of the changes across both guidance modes. For example, P8 expressed difficulty distinguishing discrete sounds: *"It was hard to remember what the tones meant and what they sounded like"*. (P8) Similarly, P9 highlights the same for continuous: *"I liked the [continuous]. But then I kind of forgot... I either knew I was going better or I was going worse. But it seemed fine."* (P9). Hearing them relative to each other aided in understanding the assigned meanings: *"Only when I heard it relative to the other one, I was like, well, this sounds more happy or sounds more complete than the last one."* (P2). This eventually led them to ignore the guidance and go with their gut: *"I was worried that [the CPAP mask] was breaking, and obviously, I forgot what this sound meant. So then that was like a little mind game and I was like, well, I guess I just keep going."* (P9).

4.3 Envisioned Ideal Experience

Participants reflected on using non-verbal audio *vs* verbal cues to indicate task progress, highlighting preference for short verbal feedback due to the ease in remembering what the guidance meant: *"If I did wrong and say just "wrong". And if I do it right, then say "great". Descriptive feedback might be better than the tones."* (P4). Participants, however, also highlighted that full verbal instructions could be incomprehensible due to unfamiliar vocabulary: *"In some ways, if it gave me instructions, I might have been very confused because I don't know the terminology for anything. [...] It's nice to be able to deal with the logic of the physical thing itself. [...] A spoken sound, but don't use the vocabulary around any of the parts – just how I'm performing relative to the task order."* (P2). Participants expressed the need for confirmatory feedback on task completion: *"I felt like I had put together everything that was put in front of me, but the tone itself didn't clearly indicate to me a completeness."* (P5).

5 DESIGNING SITUATED CONVERSATIONAL AGENT SYSTEMS FOR TASK GUIDANCE

Human conversations are not always strict text or speech-based exchanges, but are multimodal between interlocutors, exchanging visual and auditory signals. This paper explored the concept of situated multimodal CAs for task guidance, wherein the Wizard played the role of a novel *video-audio situated CA* that employed the contextual vision from the input task action video and generated non-verbal auditory feedback (Figure 1). We then explored the elements of a futuristic situated CA system (Figure 5) to explore the design space of such CAs. Based on this early system design exploration, we find that a balance between use-case specificity, design complexity, and computational cost should be considered. For example, CAs specific to a use-case may have low design complexity compared to a generalized CA, however, optimizing for computational cost may lead to brittleness to task applications. We expand on the dimensions for design complexity and computational cost below.

5.1 Guidance Design: Role of Non-verbal Cues

Design complexity is influenced by the need to balance cognitive load added by the guidance itself with the task context needs, *i.e.,* high performance and quality of task completion for physical task scenarios (*e.g.,* workers in factories). We found participants preferred verbal cues for indicating progress through short confirmatory phrases such as "that's right" or "wrong", but preferred non-verbal cues for guidance as the textual instructions could contain complex jargon. Despite increasing the design complexity, coupling such short verbal cues with non-verbal guidance might be warranted. Such clustered feedback frequency is qualitatively analogous to our discrete mode, perceived to be cognitively less demanding. However, adding even simple phrases might increase cognitive demand and interfere with task performance [16]. Further investigation is needed to validate for this context. We would also need a new taxonomy of non-verbal cues for guidance and feedback. A few design dimensions include timing, tone differentiability, semantic differentiability, and mapping task characteristics to the guidance to handle task complexity [29] and task performance [30].

5.2 System Components

Computational cost is determined by the number and complexity of the individual components in a *video-audio situated CA* system. Traditional speech-based CAs have five main components: automatic speech recognition (ASR), spoken language understanding, dialogue management, natural language generation, and text-to-speech synthesis [23]. However, given the physical task context, the envisioned CA system would include additional functions beyond the initial action recognition and the final auditory feedback generation, such as emotion perception and environment sensing (Figure 5). Our initial exploration included performing action recognition by contextual scene analysis using semantic segmentation and mapping. This allows for visual scenes to be broken down as text (*e.g.,* "hands holding a bottle"), which can then function in a

Figure 5: Vision of a situated multimodal CA system, where facial expressions could be analyzed for user state detection while the task action is recognized in real-time. The right grid shows our early exploration of using semantic segmentation techniques. These systems could also be useful for larger-scale physical tasks beyond our current focus on tabletop tasks.

similar conversational exchange as a traditional CA where the interlocutors would use multimodal input-output signals. Furthermore, intermediate functions of the system can follow analogous steps of traditional systems such as dialogue management, which would primarily be affected by the dynamic and resolution of the interaction rather than solely on the input/output data format. While we studied this space using tabletop assembly tasks, we can imagine the applicability of these systems for larger physical assembly tasks as well.

6 CONCLUSION

This work explores *video-audio situated CAs*, a novel multimodal task-oriented dialog system for physical task guidance. Through a preliminary Wizard-of-Oz study (N=8), we explored the role of audio cues in these systems. While most participants experienced non-verbal cues as helpful, including lower mental demand for the discrete mode, the sounds were abstract. Through further careful design, we plan to create a taxonomy of non-verbal guidance cues that serve diverse guidance needs. We also explored the system design through a simple implementation of action recognition. We make a case for extending these systems with multi-sensory multimodal input (*e.g.,* video and environment sensors) and output (*e.g.,* audio and video) modalities, to support the presented vision of these systems.

ACKNOWLEDGMENTS

We would like to thank all our participants for their time. Special thanks go to Mike Kuniavsky, Mirjana Spasojevic, Alexandria Pabst, and Wendy Ju for their support and guidance.

REFERENCES

[1] 2024. Hello GPT-4o. https://openai.com/index/hello-gpt-4o/.
[2] 2024. Project Astra. https://deepmind.google/technologies/gemini/project-astra/.
[3] Patricia Baggett and Andrzej Ehrenfeucht. 1991. Building physical and mental models in assembly tasks. *International Journal of Industrial Ergonomics* 7, 3 (1991), 217–227.
[4] Francesco N Biondi, Angela Cacanindin, Caitlyn Douglas, and Joel Cort. 2021. Overloaded and at work: Investigating the effect of cognitive workload on assembly task performance. *Human factors* 63, 5 (2021), 813–820.
[5] Meera M Blattner, Denise A Sumikawa, and Robert M Greenberg. 1989. Earcons and icons: Their structure and common design principles. *Human–Computer Interaction* 4, 1 (1989), 11–44.
[6] Stephen A Brewster. 1997. Using non-speech sound to overcome information overload. *Displays* 17, 3-4 (1997), 179–189.
[7] Paul A Crook, Shivani Poddar, Ankita De, Semir Shafi, David Whitney, Alborz Geramifard, and Rajen Subba. 2019. SIMMC: Situated Interactive Multi-Modal Conversational Data Collection And Evaluation Platform. *arXiv preprint arXiv:1911.02690* (2019).
[8] Nils Dahlbäck, Arne Jönsson, and Lars Ahrenberg. 1993. Wizard of Oz studies—why and how. *Knowledge-based systems* 6, 4 (1993), 258–266. Publisher: Elsevier.
[9] Abhraneil Dam, Arsh Siddiqui, Charles Leclercq, and Myounghoon Jeon. 2024. Taxonomy and definition of audio augmented reality (AAR): A grounded theory study. *International Journal of Human-Computer Studies* 182 (2024), 103179.
[10] Mira El Kamali, Leonardo Angelini, Denis Lalanne, Omar Abou Khaled, and Elena Mugellini. 2020. Multimodal conversational agent for older adults' behavioral change. In *International Conference on Multimodal Interaction*. 270–274.
[11] Marina Robertovna Gozalova, Magomed Gazilovich Gazilov, Olga Victorovna Kobeleva, Maria Igorevna Seredina, and Elena Sergeevna Loseva. 2016. Non-verbal communication in the modern world. *Mediterranean Journal of Social Sciences* 7, 4 (2016), 553–553. https://doi.org/10.36941/mjss
[12] Renan Guarese, Emma Pretty, Aidan Renata, Deb Polson, and Fabio Zambetta. 2024. Exploring audio interfaces for vertical guidance in augmented reality via hand-based feedback. *IEEE Transactions on Visualization and Computer Graphics* (2024).
[13] Marti A Hearst, J Allen, C Guinn, and Eric Horvitz. 1999. Mixed-initiative interaction: Trends and controversies. *IEEE Intelligent Systems* 14, 5 (1999), 14–23.
[14] Eric Horvitz. 1999. Principles of Mixed-Initiative User Interfaces. In *CHI '99: Proceedings of the SIGCHI conference on Human Factors in Computing Systems*. 159 – 166.
[15] Samuel Kernan Freire, Mina Foosherian, Chaofan Wang, and Evangelos Niforatos. 2023. Harnessing large language models for cognitive assistants in factories. In *Proceedings of the 5th International Conference on Conversational User Interfaces*. 1–6.
[16] Roberta L Klatzky, James R Marston, Nicholas A Giudice, Reginald G Golledge, and Jack M Loomis. 2006. Cognitive load of navigating without vision when guided by virtual sound versus spatial language. *Journal of experimental psychology: Applied* 12, 4 (2006), 223.
[17] Satwik Kottur, Seungwhan Moon, Alborz Geramifard, and Babak Damavandi. 2021. SIMMC 2.0: A Task-oriented Dialog Dataset for Immersive Multimodal Conversations. *EMNLP 2021 - 2021 Conference on Empirical Methods in Natural Language Processing, Proceedings* (2021), 4903–4912. https://doi.org/10.18653/V1/2021.EMNLP-MAIN.401
[18] Sheetal Kusal, Shruti Patil, Jyoti Choudrie, Ketan Kotecha, Sashikala Mishra, and Ajith Abraham. 2022. AI-based conversational agents: A scoping review from technologies to future directions. *IEEE Access* 10 (2022), 92337–92356.
[19] Ze-Hao Lai, Wenjin Tao, Ming C Leu, and Zhaozheng Yin. 2020. Smart augmented reality instructional system for mechanical assembly towards worker-centered

intelligent manufacturing. *Journal of Manufacturing Systems* 55 (2020), 69–81.
[20] Steven M LaValle. 2023. *Virtual Reality*. Cambridge University Press.
[21] Haeju Lee, Oh Joon Kwon, Yunseon Choi, Minho Park, Ran Han, Yoonhyung Kim, Jinhyeon Kim, Youngjune Lee, Haebin Shin, Kangwook Lee, et al. 2022. Learning to embed multi-modal contexts for situated conversational agents. In *Findings of the Association for Computational Linguistics: NAACL 2022*. 813–830.
[22] Qing Lin and Youngjoon Han. 2014. A context-aware-based audio guidance system for blind people using a multimodal profile model. *Sensors* 14, 10 (2014), 18670–18700.
[23] Ramón López-Cózar, Zoraida Callejas, Gonzalo Espejo, and David Griol. 2011. Enhancement of conversational agents by means of multimodal interaction. In *Conversational agents and natural language interaction: techniques and effective practices*. IGI Global, 223–252.
[24] Manohar Madan, Tom Bramorski, and RP Sundarraj. 1995. The effects of grouping parts of ready-to-assemble products onassembly time: an experimental study. *International Journal of Operations & Production Management* 15, 3 (1995), 39–49.
[25] Fatik Baran Mandal. 2014. Nonverbal communication in humans. *Journal of human behavior in the social environment* 24, 4 (2014), 417–421. https://doi.org/10.1080/10911359.2013.831288
[26] Nikolaos Mavridis. 2007. *Grounded situation models for situated conversational assistants*. Ph. D. Dissertation. Massachusetts Institute of Technology.
[27] Seungwhan Moon, Satwik Kottur, Paul A. Crook, Ankita De, Shivani Poddar, Theodore Levin, David Whitney, Daniel Difranco, Ahmad Beirami, Eunjoon Cho, Rajen Subba, and Alborz Geramifard. 2020. Situated and Interactive Multimodal Conversations. *COLING 2020 - 28th International Conference on Computational Linguistics, Proceedings of the Conference* (2020), 1103–1121. https://doi.org/10.18653/V1/2020.COLING-MAIN.96
[28] Tomi Nukarinen, Roope Raisamo, Ahmed Farooq, Grigori Evreinov, and Veikko Surakka. 2014. Effects of directional haptic and non-speech audio cues in a cognitively demanding navigation task. In *Proceedings of the 8th Nordic Conference on Human-Computer Interaction: Fun, Fast, Foundational*. 61–64.
[29] Rafael Radkowski, Jordan Herrema, and James Oliver. 2015. Augmented reality-based manual assembly support with visual features for different degrees of difficulty. *International Journal of Human-Computer Interaction* 31, 5 (2015), 337–349.
[30] Miles Richardson, Gary Jones, and Mark Torrance. 2004. Identifying the task variables that influence perceived object assembly complexity. *Ergonomics* 47, 9 (2004), 945–964.
[31] Anna Rouben and Loren Terveen. 2007. Speech and non-speech audio: Navigational information and cognitive load. In *International Conference on Auditory Display*. http://hdl.handle.net/1853/50039
[32] Manaswi Saha, Wendy Ju, Mike Kuniavsky, and David Goedicke. 2023. Audio AR: An Introduction. https://medium.com/labs-notebook/audio-ar-an-introduction-698661405ff4. Accessed on 24 January 2024.
[33] Jason Sterkenburg, Steven Landry, Shabnam FakhrHosseini, and Myounghoon Jeon. 2023. In-vehicle air gesture design: impacts of display modality and control orientation. *Journal on Multimodal User Interfaces* 17, 4 (2023), 215–230.
[34] Anirudh Sundar and Larry Heck. 2022. Multimodal Conversational AI: A Survey of Datasets and Approaches. *arXiv preprint arXiv:2205.06907* (2022).
[35] Anna Syberfeldt, Oscar Danielsson, Magnus Holm, and Lihui Wang. 2015. Visual assembling guidance using augmented reality. *Procedia Manufacturing* 1 (2015), 98–109.
[36] Keishi Tainaka, Yuichiro Fujimoto, Masayuki Kanbara, Hirokazu Kato, Atsunori Moteki, Kensuke Kuraki, Kazuki Osamura, Toshiyuki Yoshitake, and Toshiyuki Fukuoka. 2020. Guideline and tool for designing an assembly task support system using augmented reality. In *2020 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*. IEEE, 486–497.
[37] William Thompson. 2007. Situated Conversational Agents. In *National Conference on Artificial Intelligence*, Vol. 22. Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999, 1908.
[38] Barbara Tropschuh, Sina Niehues, and Gunther Reinhart. 2021. Measuring physical and mental strain during manual assembly tasks. *Procedia CIRP* 104 (2021), 968–974.
[39] Yunyi Yang, Yunhao Li, and Xiaojun Quan. 2021. Ubar: Towards fully end-to-end task-oriented dialog system with gpt-2. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 35. 14230–14238.
[40] Victor W Zue and James R Glass. 2000. Conversational interfaces: Advances and challenges. *Proc. IEEE* 88, 8 (2000), 1166–1180.