Designing Non-Humanoid Virtual Assistants for Task-Oriented AR Environments

Bettina Schlager* Steven K. Feiner†

Department of Computer Science
Columbia University

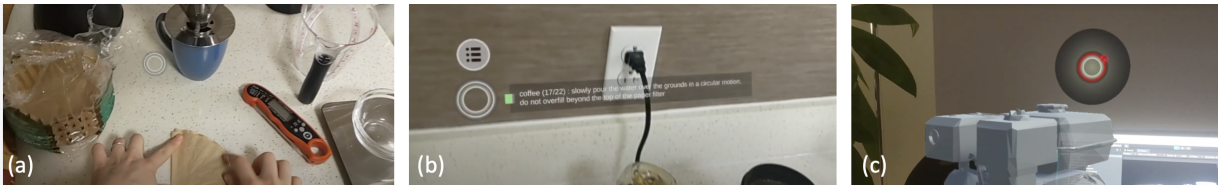

Figure 1: Our non-humanoid ECA in two testbed environments: (a–b) recipe guidance and (c) equipment maintenance. (a) In an 'idle' state, the agent consists of two concentric rings. (b) When the user looks at the agent, it displays task or notification details. (c) To notify the user about new information, such as a warning, a halo [3] attracts the user's attention until they look at the agent.

Abstract

In task-oriented Augmented Reality (AR), humanoid Embodied Conversational Agents can enhance the feeling of social presence and reduce mental workload. Yet, such agents can also introduce social biases and lead to distractions. This presents a challenge for AR applications that require the user to concentrate mainly on a task environment. To address this, we introduce a non-humanoid virtual assistant designed for minimal visual intrusion in AR. Our approach aims to enhance a user's focus on the tasks they need to perform. We explain our design choices based on previously published guidelines and describe our prototype implemented for an optical–see-through headset.

Index Terms: Computing methodologies—Computer graphics—Graphics systems and interfaces—Mixed / augmented reality; Human-centered computing—Human–computer interaction (HCI)—Interaction paradigms—Mixed / augmented reality

1 Introduction

Embodied Conversational Agents (ECAs) find widespread use in Augmented Reality (AR) and Virtual Reality (VR) in healthcare, training, and education. An ECA often acts as an assistive presence [11]. If shaped like a human, then facial expressions, body movements, or skin color could lead to social biases or the uncanny valley effect [7]. Many researchers have explored alternative forms of representation, such as animals, objects, or abstract shapes, to avoid these issues.

In human–computer interaction, anthropomorphism describes the use of human-like characteristics in system design to make them more relatable [5]. Utilizing this design approach, agents embodied by simple geometric shapes are potentially domain-independent and less distracting. This is a key aspect in task-oriented AR environments, where users primarily interact with their task. Yet, they are underrepresented in studies about immersive virtual agents [11].

We introduce a non-humanoid assistant, illustrated in Fig. 1, designed for task-oriented AR applications. Our design objective is to minimize distracting agent behavior during task assistance. We accomplish this here by exploring common design parameters used for virtual agents—embodiment, behavior, and voice—and how to adapt them towards our objective to effectively communicate assistive information to the user [7, 14].

*e-mail: b.schlager@columbia.edu
†e-mail:feiner@cs.columbia.edu

2 Related Work

In the evolution of ECAs, Bates [2] used emotionally expressive spheroidal agents to interact with a user's avatar in a 2D animated world. Later, Apple Inc. introduced Siri, a color-changing sphere that transforms during vocal interactions. In immersive environments, non-humanoid conversational agents are predominantly embodied using anthropomorphic objects or robots [11]. For instance, Fitton et al. [6] compared a humanoid agent to an animated yellow sphere with eyes, controlled using keyframe animations simulating idle and blinking behaviors. Wang et al. [13] conducted a study in which participants preferred a miniature humanoid agent over the virtual representation of the 2018 version of Amazon Echo, a cylinder. Their participants perceived the latter as inanimate despite its conversational abilities.

3 Design Considerations

A wide range of literature provides design principles for ECAs and their components. For example, the XAIR framework for trustworthy and effective Explainable AI in AR [14] emphasizes timing, content, and modality of AI explanations, developed with and validated by designers and end-users. Other research [7] indicates that multimodal interactions, such as combining gestures and speech, are essential for reducing communication errors in ECAs. Based on these design guidelines and our objective to optimize for assistance while users engage with their task environment, we decided that our agent should have the following characteristics:

Non-intrusive Behavior (NB) The audiovisual appearance and behavior should not cause distractions, allowing users to concentrate on their task space; **Natural Interactions (NI)** Communications with the agent should employ dialogue patterns that resemble natural conversations; **Just-In-Time Guidance (JIT)** The agent should provide instructions when needed. By offering timely and relevant task information, the agent assists users in completing tasks effectively; **Error Assistance (EA)** In case of errors, the agent should intervene in time, offering assistance for corrective actions; **Directions (D)** The agent should direct the user's attention to specific points of interest within the task environment; **Persistent Availability (PA)** The user should have on-demand access to the information the agent

holds independent of the time and the user's location within the task space; **Spatial Awareness (SA)** The agent should be aware of the physical and virtual spatial environment of the user.

4 Implementation

We implemented a prototype of our agent, shown in Fig. 1, using Unity 2020.3.25f [1] and Microsoft Mixed Reality Toolkit (MRTK) 2.7 [10], running on a Microsoft HoloLens 2 headset [9]. Similar to Gilles et al. [7], we divide the agent's properties into two groups: communication/interaction properties and physical properties.

4.1 Communication/Interaction Properties

Integrating verbal and nonverbal communication cues is essential for aligning the user's perception of the system with its actual state in an ECA, to address characteristic **NI**. Regarding the conversational aspect, the user can communicate with the agent by first using a prespecified invocation phrase to get its attention, followed by an utterance. For our prototype, we used GPT-3 [12] to generate a text response to the user's utterance and the MRTK text-to-speech feature to create spatialized audio for the agent's response.

Our agent is visualized using two concentric rings, billboarded to always face the user, shown in Fig. 1(a). We chose this design for its ability to depict free-floating motion in the 3D environment. Compared to other simple geometric forms, such as circles of the same diameter as the rings, the rings occlude less of the environment, while providing room for embedding anthropomorphic elements that enhance the communication between the agent and the user. While processing a user's utterance, the rings change into arcs and rotate, symbolizing a 'thinking' state, shown in Fig. 2(a). Additionally, Fig. 2(b) depicts two lines above the outer ring that function as 'eyes', expressing mutual gaze, which appear only when the user looks directly at the agent determined through eye tracking. As the agent speaks, its inner ring expands and contracts in sync with the amplitude of the speech data, inspired by human lip movements. The user can say 'stop' at any time, and the agent will stop talking.

Fig. 1c illustrates the use of a halo [3], to notify the user about newly emerging information to address characteristics **JIT** and **EA**. Upon initiating a notification, a halo originates from the agent and gradually expands until the user looks at the agent. The user is compelled to pay attention to the agent over time to avoid having the halo occupy a significant portion of their field of view (FOV). If the notification is a 'warning', (indicating a hazard) the outer ring is shaded 'red'.

4.2 Physical Properties

Addressing characteristics **PA** and **SA**, the agent stays within the FOV of the display. The agent avoids staying inside physical structures, such as walls or furniture, for comfortable engagement. We use the MRTK spatial awareness feature to get the bounds of the physical environment. If the user is not satisfied with the agent's current position, they can grab it, using a pinch gesture, and place it temporarily in any other 3D position. Regarding **D**, the agent holds current task information, such as instructions or notifications.

We use eye gaze, similar to McNamara et al. [8], to understand the user's attention and minimize visual clutter. Notifications or task information positioned at the agent are shown to the user only if they look at the agent, as seen in Fig. 1b. We incorporated Bell et al.'s view management methods for minimal interference with the physical task environment while the user is engaged [4]. This method enables us to optimally place the agent within the user's FOV without blocking the user's hands while engaged in physical activities within the environment, addressing the **NB** characteristic.

5 Conclusions and Future Work

We designed and implemented a prototype non-humanoid virtual assistant for just-in-time task guidance in AR. For future work, we aim to explore the performance of our design, involving comparing it with humanoid agents in terms of task completion speed and accuracy. We also plan to assess the cognitive load experienced by users and their level of trust in the system when interacting with our non-humanoid agent versus a conventional humanoid agent.

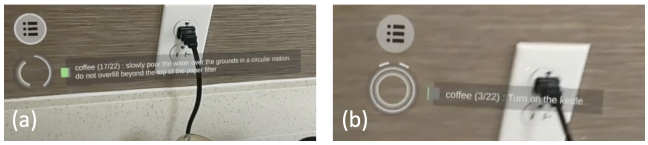

Figure 2: (a) While processing a user's utterance, two rotating arcs symbolize the agent's 'thinking' state. (b) Two lines above the outer ring function as eyes, expressing mutual gaze, which appear only when the user looks directly at the agent.

Acknowledgments

This work was supported in part by DARPA Award HR001122C0007 to Kitware, Inc. We thank Zhou Yu, Max Chen, and Derek Ahmed for developing the natural language processing components to enable a spoken dialogue between the virtual agent and the user.

References

[1] Unity. `https://unity.com/`, 2023. [Accessed 17-Dec-2023].

[2] J. Bates. The role of emotion in believable agents. *CACM*, 37(7):122–125, 1994. doi: 10.1145/176789.176803

[3] P. Baudisch and R. Rosenholtz. Halo: A technique for visualizing off-screen objects. In *Proc. CHI*, p. 481–488, 2003. doi: 10.1145/642611.642695

[4] B. Bell and S. Feiner. Representing and processing screen space in augmented reality. In M. Haller, M. Billinghurst, and B. Thomas, eds., *Emerging technologies of augmented reality: Interfaces and design*, pp. 110–137. IGI Global, 2007. doi: 10.4018/978-1-59904-066-0.ch006

[5] B. R. Duffy. Anthropomorphism and the social robot. *Robotics and Autonomous Systems*, 42(3):177–190, 2003. Socially Interactive Robots. doi: 10.1016/S0921-8890(02)00374-3

[6] I. S. Fitton, D. J. Finnegan, and M. J. Proulx. Immersive virtual environments and embodied agents for e-learning applications. *PeerJ Computer Science*, 6:e315, 2020. doi: 10.7717/peerj-cs.315

[7] M. Gilles and E. Bevacqua. A review of virtual assistants' characteristics: Recommendations for designing an optimal human–machine cooperation. *J. Comp. and Inf. Sci. in Eng.*, 22(5):050904, 03 2022. doi: 10.1115/1.4053369

[8] A. McNamara and C. Kabeerdoss. Mobile augmented reality: Placing labels based on gaze position. In *ISMAR-Adjunct*, pp. 36–37, 2016. doi: 10.1109/ISMAR-Adjunct.2016.0033

[9] Microsoft Corporation. Microsoft HoloLens 2. `https://www.microsoft.com/en-us/hololens/`, 2023. [Accessed 17-Dec-2023].

[10] Microsoft Corporation. Mixed reality toolkit (MRTK). `https://github.com/microsoft/MixedRealityToolkit-Unity`, 2023. [Accessed 17-Dec-2023].

[11] N. Norouzi, K. Kim, G. Bruder, A. Erickson, Z. Choudhary, Y. Li, and G. Welch. A systematic literature review of embodied augmented reality agents in head-mounted display environments. In *Proc. ICAT and Eurographics Symp. on VEs*, 2020. doi: 10.2312/egve.20201264

[12] OpenAI. GPT-3. `https://openai.com/blog/gpt-3-apps`, 2023. [Accessed 17-Dec-2023].

[13] I. Wang, J. Smith, and J. Ruiz. Exploring virtual agents for augmented reality. In *Proc. CHI*, pp. 1–12, 2019. doi: 10.1145/3290605.3300511

[14] X. Xu, A. Yu, T. R. Jonker, K. Todi, F. Lu, X. Qian, J. M. Evangelista Belo, T. Wang, M. Li, A. Mun, T.-Y. Wu, J. Shen, T. Zhang, N. Kokhlikyan, F. Wang, P. Sorenson, S. Kim, and H. Benko. XAIR: A framework of explainable AI in augmented reality. In *Proc. CHI*, pp. 1–30, 2023. doi: 10.1145/3544548.3581500