Shared Reality: Spatial Intelligence in Intuitive User Interfaces

Tom Stocky and Justine Cassell
MIT Media Lab
Cambridge, MA, USA
{tstocky, justine}@media.mit.edu

Abstract
In this paper, we describe an interface that demonstrates spatial intelligence. This interface, an embodied conversational kiosk, builds on research in embodied conversational agents (ECAs) and on information displays in mixed reality and kiosk format. ECAs leverage people's abilities to coordinate information displayed in multiple modalities, particularly information conveyed in speech and gesture. Mixed reality depends on users' interactions with everyday objects that are enhanced with computational overlays. We describe an implementation, MACK (Media lab Autonomous Conversational Kiosk), an ECA who can answer questions about and give directions to the MIT Media Lab's various research groups, projects and people. MACK uses a combination of speech, gesture, and indications on a normal paper map that users place on a table between themselves and MACK. Research issues involve users' differential attention to hand gestures, speech and the map, and how reference using these modalities can be fused in input and generation.

Keywords
Public information kiosks, interaction techniques, multimodal systems, embodied conversational agents.

INTRODUCTION
The old-fashioned information booths in railway stations, department stores and museums had one significant advantage over today's information kiosk: staff members could rely on the physical space shared with a visitor in order to give directions, describe travel and spatialize relationships among places and things. Railway personnel pointed out the proper train platform; department store greeters unfolded store maps to give directions to shoppers; and museum staff illustrated the size of the dinosaurs in the great hall. This paper describes an intelligent kiosk that integrates the face-to-face strengths of information booths with the self-sufficiency and accuracy of information access stations by incorporating an Embodied Conversational Agent (ECA) into an information kiosk.

IUI'02, January 13-16, 2002, San Francisco, California, USA.

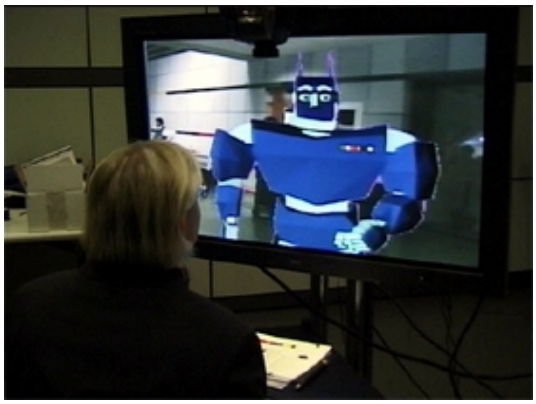

Figure 1: User interacting with MACK.

ECAs have served a wide variety of purposes, from tutoring to sales. Examining their past use reveals some of the strengths of such systems. For example, in the tutoring realm, ECAs have served as animated pedagogical agents that taught students procedural tasks in simulated environments [8]. Research indicates that such animated agents provide key benefits that enhance learning environments. One such benefit is that embodied agents serve as valuable navigational guides that can direct students and show them how to get around. Virtual 3D learning environments represent a further advance in navigational guidance by helping students develop spatial models of the subject matter [6]. ECA Kiosks were conceived with this in mind, as a logical extension of ECAs as navigational guides in virtual worlds. However, instead of immersing the ECA in a 3D virtual world, ECA Kiosks immerse both system and user in the *actual physical space*, allowing them to interact within the shared physical and informational reality they are referencing.

Research indicates that public information kiosks are useful and effective interfaces. They have been shown to increase user acceptance of the online world, in that they serve a wide range of individuals. Knowledge transfer is also improved, as kiosk users have demonstrated that they gain new information and tend to use the system repeatedly after initial interactions. And kiosks increase business utility by increasing the likelihood of purchase and reducing the time needed for physical staff to provide advice [9].

However, current kiosks have been limited in interaction techniques, requiring literacy on the part of users, and the use of one's hands to type or choose information. Replacing text and graphics with an ECA may result in

systems that are more flexible, allowing for a wider diversity in users. ECAs allow for hands-free multimodal input and output (speech and gesture), which produces a more natural, more intuitive interaction [1]. These communication protocols come without need for user training, as all users have these skills and use them daily [2]. Natural language and gesture take full advantage of the shared environment, which creates a spatial bridge between the user and the agent.

THE SYSTEM: MACK

In designing and creating MACK, an ECA Kiosk, we had three primary requirements:

1) Real-time multimodal input as a basis for natural face-to-face interaction.
2) Coordinated natural language and gesture generation.
3) The ability to reference physical reality. Hence, the system must be aware of its location and orientation, as well as the layout of the physical building.

MACK currently employs three types of input: (1) speech recognition via MIT LCS' SpeechBuilder technology [5], (2) user presence using a pressure-sensing chair mat, and (3) deictic map input via a paper map atop a table with an embedded Wacom tablet. These three are merged following Johnston's finite-state transducer approach [7]. For example, a user can say, "Tell me about this" while pointing to a specific research group on the map, and MACK will respond with information about that group.

MACK uses multimodal output as well, with (1) speech synthesis using the Microsoft Whistler Text-to-Speech (TTS) engine, (2) an LCD projector output directed at the physical map to highlight areas and draw paths between points, and (3) on-screen graphical output including synchronized head, arm and eye movements. BEAT [3] is responsible for the generation of appropriate speech with intonation, hand gesture, and head and eye movements. Input to BEAT is typed text, generated using templates based on information from the database.

From the user's perspective, MACK is a life-sized on-screen blue robot seemingly immersed in their shared physical environment. This is achieved with a video mixer and camera mounted atop the plasma screen display. On the screen behind MACK appears a direct video feed of the physical background. Since MACK is aware of his physical location and orientation, he is able to say things like, "It's right behind me," and point back with his thumb.

To define MACK's coordination of speech, hand gestures and map-annotations, we studied natural human-to-human direction-giving. Eleven subjects were told to find their way to two distinct locations in the Media Lab by standing by the elevators, where there was a map of the building, and asking for help from passersby, requesting clarification – e.g., "I'm not sure I understand …" – after the second set of directions.

Similar to findings by Tversky [10], direction-givers employed three methods for direction-giving: (1) relative descriptions – e.g., "Do you know where the coffee machine is? It's next to that." or "It's on the third floor." – (2) explanations with deictic gestures, and (3) map-based route planning. These three methods were used progressively, to disambiguate misunderstanding. That is, when possible, the direction-giver provided a relative description, based on either an assumed or established context. If the direction-receiver was unsatisfied, the route was explained with speech and deictic gestures. If the receiver remained unsatisfied, the route was then defined with speech and map-based gesture. We are in the process of translating these results into a computational module.

FUTURE WORK

MACK will soon be able to give directions based on prior interaction and shared context. At that point we will collect data on human interaction with the system to determine if spatial intelligence of this sort results in improved user performance on route finding.

REFERENCES

[1] Cassell, J., T. Bickmore, M. Billinghurst, L. Campbell, K. Chang, H. Vilhjálmsson, and H. Yan. "Embodiment in Conversational Interfaces: Rea." Proc. *CHI '99*, Pittsburgh, PA (1999): 520-527.

[2] Cassell, J., T. Bickmore, L. Campbell, H. Vilhjálmsson, and H. Yan. "More Than Just a Pretty Face: Conversational Protocols and the Affordances of Embodiment." *Knowledge-Based Systems* 14 (2001): 55-64.

[3] Cassell, J., H. Vilhjálmsson, et al. "BEAT: The Behavior Expression Animation Toolkit." Proc. *SIGGRAPH '01*. Los Angeles, CA (August 2001).

[4] Chang, J. "Action Scheduling in Humanoid Conversational Agents." M.S. Thesis in Electrical Engineering and Computer Science. Cambridge, MA: MIT (1998).

[5] Glass, J. and E. Weinstein. "SpeechBuilder: Facilitating Spoken Dialogue System Development." Proc. *EuroSpeech '01*. Aarlborg, Denmark (September 2001).

[6] Johnson, W., J. Rickel, and J. Lester. "Animated pedagogical agents: Face-to-face interaction in interactive learning environments." *International Journal of Artificial Intelligence in Education* 11 (2000): 47-78.

[7] Johnston, M. and S. Bangalore. "Finite-State Multimodal Parsing and Understanding." Proc. *COLING-2000*, Saarbruecken, Germany (August 2000).

[8] Rickel, J., N. Lesh, C. Rich, C.L. Sidner, and A. Gertner. "Building a Bridge between Intelligent Tutoring and Collaborative Dialogue Systems." Proc. *Tenth International Conference on AI in Education* (2001): 592-594. IOS Press.

[9] Steiger, P and B. Ansel Suter. "MINELLI – Experiences with an Interactive Information Kiosk for Casual Users." Proc. *UBILAB '94*, Zurich (1994): 124-133.

[10] Tversky, B. "Spatial schemas in depictions." In *Spatial Schemas and Abstract Thought*. Ed. M. Gattis. Cambridge, MA: MIT Press, 1999.