In *Proceedings of IMAGINA'02*, Monte Carlo, Jan 12-15, 2002)

MACK: Media lab Autonomous Conversational Kiosk

Justine Cassell, Tom Stocky, Tim Bickmore, Yang Gao, Yukiko Nakano, Kimiko Ryokai, Dona Tversky, Catherine Vaucelle, Hannes Vilhjálmsson

MIT Media Lab
Cambridge, MA, 02145 USA
justine@media.mit.edu

Abstract

In this paper, we describe an embodied conversational kiosk that builds on research in embodied conversational agents (ECAs) and on information displays in mixed reality and kiosk format in order to display spatial intelligence. 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 flexible architectures for Embodied Conversational Agents that allow these modalities to be fused in input and generation.

INTRODUCTION

Research in computational linguistics, multimodal interfaces, computer graphics, and autonomous agents has led to the development of increasingly sophisticated autonomous or semi-autonomous virtual humans over the last five years.

Autonomous self-animating characters of this sort are important for use in production animation, interfaces and computer games. And increasingly their autonomy comes from underlying models of behavior and intelligence, rather than simple physical models of human motion. Intelligence also increasingly refers not just to the ability to reason, but also to "social smarts" – the ability to engage a human in an interesting, relevant conversation with appropriate speech and body behaviors. Our own research concentrates on the type of virtual human that has the social and linguistic abilities to carry on a face-to-face conversation, what we call Embodied Conversational Agents.

Embodied conversational agents may be defined as those that have the same properties as humans in face-to-face conversation, including:

- The ability to recognize and respond to verbal and non-verbal input
- The ability to generate verbal and non-verbal output.
- The use of conversational functions such as turn taking, feedback, and repair mechanisms.
- A performance model that allows negotiation of the conversational process, and contributions of new propositions to the discourse.

The last five years have also seen a continuing trend to develop, and to situate in public locations, information access systems that can deliver resources and services to the general public. These kiosk systems place interesting constraints on interface design. Kiosk systems must stand out so that they will be noticed by casual passers-by, and their purpose must be self-evident. Users of kiosk systems do not have time for lengthy training, and so interaction must be self-explanatory. User bases for public spaces tend to represent a diverse set of backgrounds, and so systems must be geared towards the least common denominator, and demonstrate the ability to recover from errors in use.

However, 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 various 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. Such an application – including the new input and output modality of a paper map – allows us to demonstrate how research into new input and output modalities can be taken into account using our existent architecture for Embodied Conversational Agents.

BACKGROUND

Human face-to-face conversation is a complex phenomenon involving understanding and synthesis across multiple modalities and time scales. Speech, intonation, gaze, and head movements function not just in parallel, but interdependently. The form of each of these modalities – a rising tone vs. a falling tone, pointing towards oneself vs. pointing towards the other – is essential to the meaning. But the co-occurrence of behaviors, how the modalities work together, and the choice of which modality to use

when are equally important. As far as co-occurrence is concerned, there is a tight synchrony among the different communicative modalities in humans. Speakers accentuate only the important words by speaking more forcefully, gesture along with the word that a gesture illustrates, and turn their eyes towards the listener when coming to the end of a thought. Meanwhile listeners nod within a few hundred milliseconds of when the speaker's gaze shifts. This synchrony is essential to the meaning of conversation. Speakers will go to great lengths to maintain it (stutterers will repeat a gesture over and over again, until they manage to utter the accompanying speech correctly) and listeners take synchrony into account in what they understand. (Readers can contrast "this is a **stellar** Imagina paper" [big head nod along with "stellar"] with "this is a **. . .** stellar Imagina paper" [big head nod during the silence]). When synchrony among different communicative modalities is destroyed, as in low bandwidth videoconferencing, satisfaction and trust in the outcome of a conversation is diminished. When synchrony among different communicative modalities is maintained, as when one manages to nod at all the right places during the Macedonian policeman's directions, despite understanding not a word, conversation comes across as successful.

As far as how the modalities work together, speech and nonverbal behaviors do not always manifest the same information, but what they convey is virtually always compatible. In many cases, different modalities serve to reinforce one another. In other cases, semantic and pragmatic attributes of the message are distributed across the modalities such that the full communicative intentions of the speaker are interpreted by combining linguistic and body language information. For example, a deictic (pointing) gesture accompanying the spoken words "that folder" may substitute for an expression that encodes all of the necessary information in the speech channel, such as "the folder on top of the stack to the left of my computer." When talking about objects for the first time, speakers tend to use gesture to indicate some aspects of the object, and speech to indicate complementary aspects of that object. When talking about the location or the shape of an object, on the other hand, they tend to redundantly use their hands and their speech to describe the thing [5]. In this sense, the semantic and pragmatic compatibility seen in the gesture-speech relationship recalls the interaction of words and graphics in multimodal presentations [8, 10, 27]. In fact, some suggest [16], that gesture and speech arise together from an underlying representation that has both visual and linguistic aspects, and so the relationship between gesture and speech is essential to the production of meaning and to its comprehension.

As far as choice of modalities is concerned, in natural conversation speakers tend to produce a gesture when introducing a new discourse entity into the dialogue. They tend to shift their gaze at the beginnings

and ends of their speaking turns. In computational pen-and-speech applications using a map, users tend to use speech and a pen together when describing spatial information about the location, number, size, orientation, or shape of an object. However, when performing general actions without any spatial component, such as printing a map, users rarely expressed themselves multimodally -- less than 1% of the time [19]. Much research has been devoted to route directions and their underlying structure and semantics. Michon and Denis examined the use of landmarks in oral direction-giving. They found that landmarks are spoken of most frequently at reorientation points along the route. Landmarks were also found useful to direction-receivers in helping them to construct mental representations of unfamiliar environments in which they are preparing to move [17]. Tversky has looked at how people use speech and maps to convey routes and has found a common structure in the use of written directions and hand-drawn maps:

> The first step is to put the listener at the point of departure. In the field, this is typically apparent to both interlocutors and need not be specified. The second step, beginning the progression, may also be implicit. The next three steps are used iteratively until the goal is reached: designate a landmark; reorient the listener; start the progression again by prescribing an action. Actions may be changes of orientation or continuations in the same direction. [26]

In our own work, we investigated the use of speech, gesture, and printed maps in 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 [25], 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 first a short relative description, based on either an assumed or established landmark. That is, for example, the direction-giver said "It's near the freight elevator", where the freight elevator served as context for the description of the location of the kitchen. If the direction-receiver was unsatisfied, the route was explained with speech and deictic gestures. In this case, the direction-giver described the route using a first-person perspective in gesture, and a second-person perspective in speech. That is, the direction-giver might say: " turn to the left, go all the way down this hallway, and then you'll see a door" but even while using the second-person pronoun "you", the direction-giver points left, then down an imaginary hall in

front of her, and then makes a flat gesture describing a door on her left. Finally, if the receiver remained unsatisfied, the route was then defined by using speech, and tracing the route on the map. However, in this case the direction-giver switched to a "survey view", without walking the person through the halls.

Synchrony and the association between different modalities allows participants in face-to-face dialogue, such as railway personnel and lost passengers, to have available information from a variety of modalities that can help them to understand what is being communicated. We approach these questions from the point of view of building communicating humanoid agents that can interact with humans -- that can, therefore, understand and produce information conveyed by the modalities of speech, intonation, facial expression and hand gesture. In order for computer systems to fully understand messages conveyed in such a manner, they must be able to collect information from a variety of channels and integrate it into a combined "meaning." While this is certainly no easy proposition, the reverse task is perhaps even more daunting. In order to *generate* appropriate multi-modal output, including speech with proper intonation and gesture, the system must be able to make decisions about how and when to distribute information across channels. In previous work, we have built a system that are able to decide where to generate gestures with respect to information structure and intonation, and what kinds of gestures to generate (iconics, metaphorics, beats, deictics) (Cassell et al, 1994), and a system that is able to decide what information should be displayed via gesture and what information should be conveyed via speech. Sometimes the department store hostess may pull out a map of the store, and sometimes she may user her hands to indicate the path to the restaurant. And so, in the current system, we need to generate either gesture, speech, or indications on a paper map.

RELATED WORK

Kiosks

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 [23].

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 [2]. These communication protocols come without need for user training, as all users have these skills and use them daily . Natural language and gesture take full advantage of the shared environment, which creates a spatial bridge between the user and the agent.

Significant research has been conducted to find ways of effectively presenting information, and ways of allowing users to interact with that information. Much research, for example, has concentrated on using touch screens to allow more intuitive interaction with bodies of information. Additional research has examined the most natural kinds of linkages between those bodies of information, and how to allow users to engage in "social navigation" – following the trails of others, or patterns generated by their own evolving interests.

The MINELLI system was created as a hypermedia public information kiosk with a touch screen interface. Rather than the standard, static, text-with-graphics content, MINELLI used short films, musical and graphical content, and interactive games to engage users and make them feel comfortable using the system [20, 23]. While MINELLI was certainly an improvement over standard kiosks, it required user training, which is not ideal for a public access system. Raisamo's Touch'n'Speak kiosk demonstrates an approach requiring no training, employing natural language input in conjunction with a touch screen. These modalities were selected with the goal of creating an intuitive interface that required no user training [20]. While a touch screen is perhaps more intuitive than a keyboard and mouse, both MINELLI and Touch'n'Speak remain limited to a primarily menu-driven process flow. Embedding an ECA into the interface addresses this limitation with the use of dialogue-based interaction, which has the added benefit of not requiring user literacy.

Others have looked at multimodal display of information, and multimodal input. Looking at the combination of text and graphics in information display, Kerbedjiev proposed a methodology for realizing communicative goals in graphics. He suggests that more appealing multimedia presentations take advantage of both natural language and graphics [13]. Such findings have paved the way for ECAs, capable of natural language and its associated nonverbal behaviors. An ECA Kiosk has a further advantage over graphics, in that it can reference actual physical objects in the real world.

Feiner and McKeown make a similar distinction between the function of pictures and words. Pictures describe physical objects more clearly, while language is more adept in conveying information about ab-

stract objects and relations. This research led to their COMET system generates text and 3D graphics on the fly [8]. Similarly, Maybury's TEXTPLAN generates multimedia explanations, tailoring these explanations based on the type of communicative act required [15]. Taking these principles to the next level means replacing graphical representations with the physical objects themselves. To do so involves immersing the interface in a shared reality, as a kiosk, and using an ECA to reference objects in the real world using natural language and gesture.

2.2 Spatial Reference

In designing interfaces capable of spatial reference, one approach is immersive virtual reality. For example, Billinghurst's intelligent medical interface was designed to allow surgeons to interact with virtual tissue and organ models, achieved through the use of continuous voice recognition coupled with gesture input via pointing and grasping of simulated medical instruments [1]. In this way, the interface immerses the user in a shared virtual reality. ECA Kiosks take this idea in a slightly different direction by obviating the need for head mounted displays, and immersing user and interface in a shared *physical* reality – the real world.

Other implementations have come closer to creating shared physical reality, through the use of devices that can accompany the user, such as PDAs. One such interface, PalmGuide, is a hand-held tour guidance system that refers to objects in the user's reality, recommending exhibits that may be of interest. The interface is primarily text-based, accented by user-specified character icons that give it an added sense of familiarity. When in the vicinity of a computer kiosk, PalmGuide is able to exchange information so that the kiosk can present information in a way that is comfortable and natural to the user [24]. ECA Kiosks take this concept of natural interaction a step further by using an animated character capable of human-like conversation with the user. And taking PalmGuide's representation of spatiality to the next level, ECA Kiosks can access, via shared reality, the same spatial references as the user.

The ability to refer to space multimodally is also addressed by OGI's QuickSet system, which allows users to reference maps through the use of pen and voice inputs. For example, a user can create an open space on a map by drawing an area and saying, "Open space." This multimodal input allows users to interact with the system in a natural and intuitive manner [18]. In designing MACK, we created a system that allows users to reference a shared physical map, similar to QuickSet in its use of pen and speech multimodal input. In addition to allowing users to reference the map in a natural way, the physical map also serves as another bridge between the real and virtual worlds.

2.3 Embodiment

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 [22]. 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 [11]. 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.

In past research, embodiment has proven its effectiveness in engaging users [14] [21], and shown a qualitative advantage over non-embodied interfaces, enabling the exchange of multiple levels of information in real time [4]. One example of this is Rea, an ECA real estate agent capable of both multimodal input and output. Users that interacted with Rea found the interface intuitive and natural, as conversation is an intrinsically human skill that requires no introduction [4]. This carries over to an ECA Kiosk such as MACK, whose interface requires no user training, and is therefore capable of success in a public access implementation.

Cambridge Research Laboratory has also explored this concept of embodiment while trying to create a better way for people to obtain information in public spaces. They began with a traditional public access kiosk, and enhanced it with an animated head and synthesized speech. The kiosk could also track the faces of passing users [28]. They deployed these kiosks in public places, and one of the foremost lessons learned was that people were attracted to an animated face that watched them [7]. CRL also report, however, that while a face-only avatar did well at attracting and entertaining people, it was not successful at conveying or interacting with content. These animated heads lack the ability to indicate spatiality through hand and arm gestures, one of the reasons we chose an ECA with a visible torso when designing MACK. As Oviatt [18] remarks,

> Although the everyperson information kiosk may be an admirable goal, its presumptions fail to acknowledge that different modes represented by the emerging technologies that recognize speech, handwriting, manual gestur-

ing, head movements, and gaze each are strikingly unique. They differ in the type of information they transmit, their functionality during communication, the way they are integrated with other modes, and in their basic suitability to be incorporated into different interface styles. None of these modes are simple analogues of one another in the sense that would be required to support simple one-to-one translation.

Architectural Requirements

As we have described in previous work [3], the construction of a computer character which can effectively participate in face-to-face conversation as described above requires a control architecture which has the following features:

- *Multi-Modal Input and Output* **–** since humans in face-to-face conversation send and receive information through gesture, intonation, and gaze as well as speech, the architecture also should support receiving and transmitting this information.

- *Real-time* –The system must allow the speaker to watch for feedback and turn requests, while the listener can send these at any time through various modalities. The architecture should be flexible enough to track these different threads of communication in ways appropriate to each thread. Different threads have different response time requirements; some, such as feedback and interruption occur on a sub-second timescale. The architecture should reflect this fact by allowing different processes to concentrate on activities at different timescales.

- *Understanding and Synthesis of Propositional and Interactional Information* – Dealing with propositional information – the content of the communication – requires building a model of user's needs and knowledge. Thus the architecture must include both a static domain knowledge base and a dynamic discourse knowledge base. Presenting propositional information requires a planning module to plan how to present multi-sentence output and manage the order of presentation of interdependent facts. Understanding interactional information – about the processes of conversation, on the other hand, entails building a model of the current state of the conversation with respect to conversational process (who is the current speaker and who is the listener, has the listener understood the speaker's contribution, and so on).

- *Conversational Function Model* – Explicitly representing conversational functions provides both modularity and a principled way to combine different modalities. Functional models influence the architecture because the core modules of the system operate exclusively on functions (rather than sen-

tences, for example), while other modules at the edges of the system translate input into functions, and functions into outputs. This also produces a symmetric architecture because the same functions and modalities are present in both input and output.

Figure 1 illustrates an architecture with these properties.

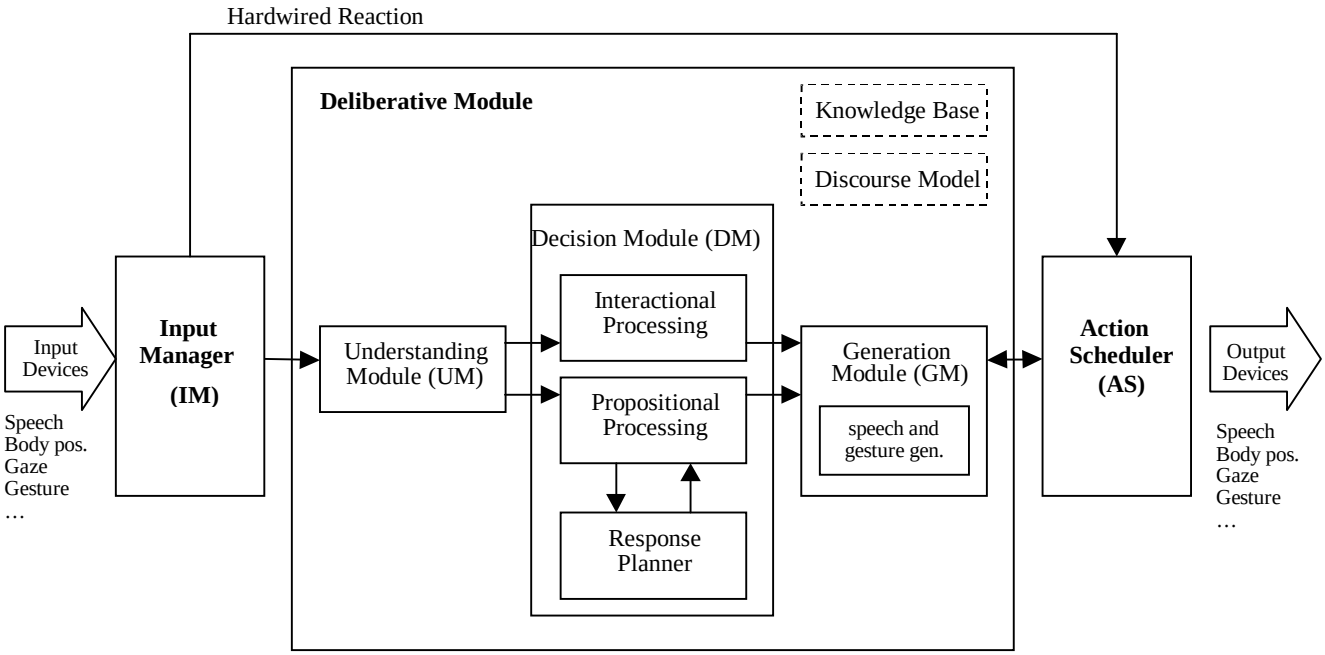

Figure 1: ECA Software Architecture

THE SYSTEM: MACK

In a system such as an information kiosk, content generation and understanding can be limited. That is, queries will usually have the same form ("where is the Gesture and Narrative Language Group") and answers need not be planned on the fly (since the answer will always be "on the third floor," regardless of the utterance which came before in the conversation). Therefore, for the purpose of MACK, we replaced the planning and complex generation module with a simple template-based sentence generator. And, rather than, as in earlier work, generating gestures and speech on the fly directly from a conceptual representation, we generated text in the Generation Module, and then sent that text to the BEAT system for nonverbal expression generation [6].

In designing and creating MACK, on the other hand, we had one additional requirement to place on this

general architecture: 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 in which it is located.

The architecture of MACK, therefore, can be simplified, as in Figure 2.

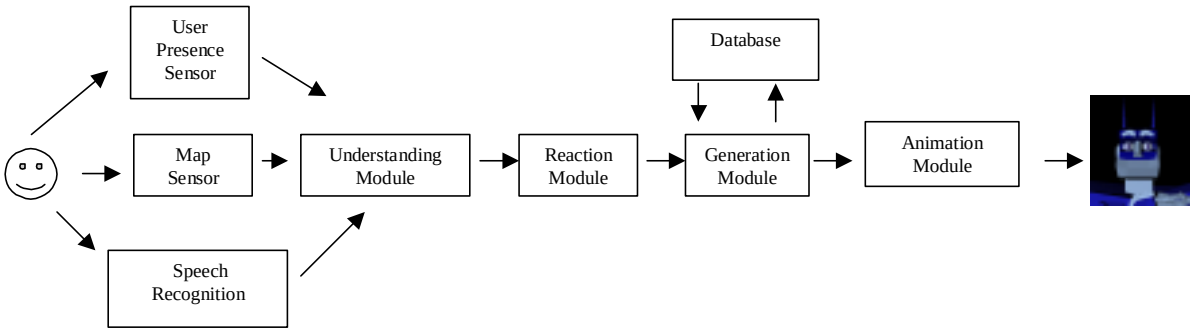

Figure 2. MACK System Architecture

The animation module of MACK plays an important role in an information kiosk where gesture plays a role in direction-giving, where a human-like interaction must be maintained, and where speech, gesture and map indications must be tightly synchronized. The animation module here is the BEAT system [6], with an architecture as shown in Figure 3. BEAT has the advantage of automatically annotating text with hand gestures, eye gaze, eyebrow movement, and intonation. The annotation is carried out in XML, through interaction with a knowledge base of the domain being discussed (in this case, a database of projects, researchers, and demos available throughout the building), and via a set of behavior generation rules. Output is scheduled in such a way that tight synchronization is maintained among modalities.

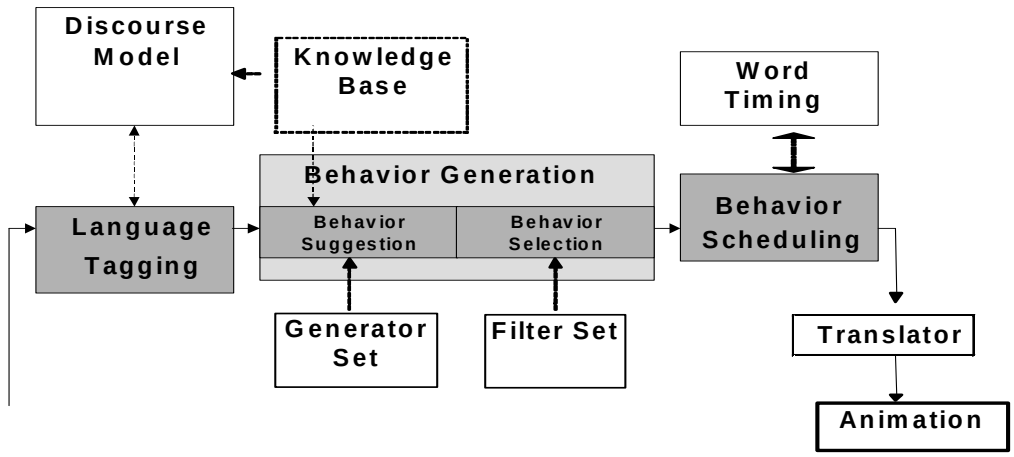

Figure 3: BEAT Text-to-Nonverbal Behavior Module

These features give us a system with functionality as follows: MACK employs three types of input: (1) grammar-based speech recognition using MIT LCS' SpeechBuilder technology [9], (2) recognition of user presence by way of a pressure-sensing chair mat, and (3) recognition of pen gestures on a paper map by way of a Wacom tablet embedded into a table that sits between the user and MACK.

All signals from the sensory inputs are aggregated and processed in the Understanding Module, which interprets meaning from each of the various inputs. More importantly, however, the understanding module also merges the separate streams of input, following Johnston's finite-state transducer approach [12]. 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.

Information from the understanding module is passed to the reaction module, which determines MACK's response. The process to determine next-action is implemented as a state machine, where MACK's outputs are represented by individual states and each input acts as an decision parameter to generate MACK's next output. For example, MACK is by default in an "idle" state; and unless the understanding module has notified that a user has become present, MACK will not be responsive to any sensory input. In this case, the user's presence (indicated by the pressure-sensitive chair mat) becomes the input that carries MACK from "Idle" to "Greeting."

After a specific reaction is determined, the Generation Module accesses the database for information necessary to formulate a response. Sentence templates allow abstract response categories from the Reaction Module to be translated into text, which is then sent to BEAT for the generation of appropriate speech with intonation, hand gesture, head and eye movements, and map indications. Multimodal output, then includes (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.

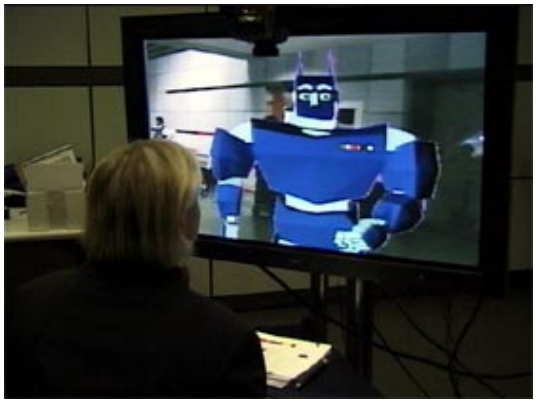

Figure 4: User interacting with MACK.

From the users' 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.

We are currently implementing a more effective direction-giving strategy, one that is drawn from the results of our empirical study. When a user asks for directions to a particular demo, MACK will begin by giving directions relative to a salient landmark, or to one that has already been discussed. If this approach is not successful, MACK will give first-person directions using hand gesture and speech. Finally, faced with a truly direction-challenged user, MACK will project a route onto the map shared with the user.

CONCLUSION

We have demonstrated MACK at several venues where approximately 200 to 300 people have used it. We observed that users' behaviors appeared natural, as though they were interacting with another person. Users acted as though MACK demonstrated agency, was trustworthy in the information he conveyed, and a good partner in their visit to the Media Lab. When MACK gave directions, for example, "room 315 is located right behind you," while pointing to the area behind the user, users turned around, and then turned back to MACK and nodded at him in thanks. MACK was also successful in engaging and entertaining users.

Demonstrating MACK also served to highlight some current technical limitations and areas for future research. Speech recognition errors were common. Due to MACK's extensive vocabulary (names of projects, names of researchers, numbers), close match errors (e.g. "vat" instead of "that") were fairly frequent. The system's limited grammar accounted for other recognition errors. For example, "Where is the Lobby?" or "How do I get to the Lobby?" would be understood, while "Where would I find the Lobby?" In the original version, before gesture and map protocols were implemented, turn-taking protocols and focus of attention were also a problem. Users had difficulty knowing when to look at the map as opposed to when they should look at MACK. We believe that the more consistent and human-like structure that we have determined will resolve this issue.

More generally, however, MACK has demonstrated that ECAs are both useful and interesting additions to public kiosks, and that our ECA model and computational architecture are adequate to the task of taking on new input and output modalities and new modes of interaction.

REFERENCES

[1] M. Billinghurst, J. Savage, P. Oppenheimer, and C. Edmond, "The Expert Surgical Assistant: An Intelligent Virtual Environment with Multimodal Input," in *Medicine Meets Virtual Reality IV: Health Care in the Information Age*. Amsterdam: IOS Press, 1996, pp. 590-607.

[2] J. Cassell, T. Bickmore, M. Billinghurst, L. Campbell, K. Chang, H. Vilhjalmsson, and H. Yan, "Embodiment in Conversational Interfaces: Rea," *Proceedings of CHI 99*, Pittsburgh, PA, 1999.

[3] J. Cassell, T. Bickmore, L. Campbell, H. Vilhjalmsson, and H. Yan, "Human Conversation as a System Framework: Designing Embodied Conversational Agents," in *Embodied Conversational Agents*, J. Cassell, J. Sullivan, S. Prevost, and E. Churchill, Eds. Cambridge, MA: MIT Press, 2000, pp. 29-63.

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

[5] J. Cassell, M. Stone, and H. Yan, "Coordination and Context-Dependence in the Generation of Embodied Conversation," *Proceedings of INLG 2000*, Mitzpe Ramon, Israel, 2000.

[6] J. Cassell, H. Vilhjálmsson, and T. Bickmore, "BEAT: The Behavior Expression Animation Toolkit," *Proceedings of SIGGRAPH 01*, Los Angeles, CA, 2001.

[7] A. D. Christian and B. L. Avery, "Speak Out and Annoy Someone: Experience with Intelligent Kiosks," *Proceedings of CHI 2000*, The Hague, Netherlands, 2000.

[8] S. K. Feiner and K. McKeown, "Automating the Generation of Coordinated Multimedia Explanations," *IEEE Computer*, vol. 24, pp. 33-41, 1991.

[9] J. Glass and E. Weinstein, "SpeechBuilder: Facilitating Spoken Dialogue System Development," *Proceedings of EuroSpeech*, Aalborg, Denmark, 2001.

[10] N. Green, G. Carenini, S. Kerpedjiev, and S. F. Roth, "A Media-Independent Content Language for Integrated Text and Graphics Generation," *Proceedings of Workshop on Content Visualization and Intermedia Representations at COLING and ACL '98*, University of Montreal, Quebec, 1998.

[11] W. L. Johnson, J. W. Rickel, and J. C. Lester, "Animated Pedagogical Agents: Face-to-Face Interaction in Interactive Learning Environments," *International Journal of Artificial Intelligence in Education*, vol. 11, pp. 47-78, 2000.

[12] M. Johnston and S. Bangalore, "Finite-state Multimodal Parsing and Understanding," *Proceedings of COLING-2000*, Saarbruecken, Germany, 2000.

[13] S. Kerpedjiev, G. Carenini, N. Green, J. Moore, and S. Roth, "Saying It in Graphics: from Intentions to Visualizations," *Proceedings of IEEE Symposium on Information Visualization*, Research Triangle Park, NC, 1998.

[14] T. Koda and P. Maes, "Agents with Faces: The Effects of Personification of Agents," *Proceedings of IEEE Robot-Human Communication '96*, Tsukuba, Japan, 1996.

[15] M. Maybury, "Planning Multimedia Explanations Using Communicative Acts," in *Readings in Intelligent User Interfaces*, M. Maybury and W. Wahlster, Eds. San Francisco: Morgan Kaufman, 1998, pp. 99-106.

[16] D. McNeill, *Hand and Mind: What Gestures Reveal about Thought*. Chicago, IL/London, UK: The University of Chicago Press, 1992.

[17] P.-E. Michon and M. Denis, "When and Why Are Visual Landmarks Used in Giving Directions," *Proceedings of Conference on Spatial Information Theory*, Morro Bay, CA, 2001.

[18] S. Oviatt and P. Cohen, "Multimodal Interfaces That Process What Comes Naturally," *Communications of the ACM*, vol. 43, pp. 45-53, 2000.

[19] S. Oviatt, A. DeAngeli, and K. Kuhn, "Integration and Synchronization of Input Modes during Multimodal Human-Computer Interaction," *Proceedings of CHI 97*, Atlanta, GA, 1997.

[20] R. Raisamo, "Evaluating Different Touch-based Interaction Techniques in a Public Information Kiosk," *Proceedings of Conference of the Computer Human Interaction Special Interest Group of the Ergonomics Society of Australia*, Charles Stuart University, 1999.

[21] B. Reeves and C. Nass, *The Media Equation: how people treat computers, televisions and new media like real people and places*. Cambridge: Cambridge University Press, 1996.

[22] J. Rickel, N. Lesh, C. Sidner, and A. Gertner, "Building a Bridge between Intelligent Tutoring and Collaborative Dialogue Systems," *Proceedings of Tenth International Conference on AI in Education*, 2001.

[23] P. Steiger and B. A. Suter, "MINELLI - Experiences with an Interactive Information Kiosk for Casual Users," *Proceedings of UBILAB*, Zurich, 1994.

[24] Y. Sumi and P. Maes, "Supporting Awareness of Shared Interests and Experiences in Community," *Proceedings of International Workshop on Awareness and the WWW, held at the ACM CSCW '00 Conference*, Philadelphia, 2000.

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

[26] B. Tversky and P. U. Lee, "Pictorial and Verbal Tools for Conveying Routes," *Proceedings of Conference on Spatial Information Theory*, Stade, Germany, 1999.

[27] W. Wahlster, E. Andre, W. Graf, and T. Rist, "Designing Illustrated Texts," *Proceedings of EACL'91*, Berlin, Germany, 1991.

[28] K. Waters and T. Levergood, "An Automatic Lip-Synchronization Algorithm for Synthetic Faces," *Proceedings of The second ACM international conference on Multimedia*, San Francisco CA, 1994.