Embodied Conversational Agents

Representation and Intelligence in User Interfaces

Justine Cassell

■ How do we decide how to represent an intelligent system in its interface, and how do we decide how the interface represents information about the world and about its own workings to a user? This article addresses these questions by examining the interaction between representation and intelligence in user interfaces. The rubric *representation* covers at least three topics in this context: (1) how a computational system is represented in its user interface, (2) how the interface conveys its representations of information and the world to human users, and (3) how the system's internal representation affects the human user's interaction with the system. I argue that each of these kinds of representation (of the system, information and the world, the interaction) is key to how users make the kind of attributions of intelligence that facilitate their interactions with intelligent systems. In this vein, it makes sense to represent a systmem as a human in those cases where social collaborative behavior is key and for the system to represent its knowledge to humans in multiple ways on multiple modalities. I demonstrate these claims by discussing issues of representation and intelligence in an embodied conversational agent—an interface in which the system is represented as a person, information is conveyed to human users by multiple modalities such as voice and hand gestures, and the internal representation is modality independent and both propositional and nonpropositional.

> Suppose that sometimes he found it impossible to tell the difference between the real men and those which had only the shape of men, and had learned by experience that there were only two ways of telling them apart: first, that these automata never answered in word or sign, except by chance, to questions put to them; and second, that though their movements were often more regular and certain than those of the wisest men, yet in many things which they would have to do to imitate us, they failed more disastrously than the greatest fools.
>
> — *Descartes, 1638*

To start with a convenient counter claim, let's examine a quote from a recent call for proposals on the topic of the disappearing computer: "[I]n this vision, the technology providing these capabilities is unobtrusively merged with real world objects and places, so that in a sense it disappears into the background, taking on a role more similar to electricity—an invisible pervasive medium." In this vision of intelligent user interfaces, there is no representation of the system and no modalities by which information is conveyed to users. One interpretation of this vision (instantiated, for example, in ubiquitous computing) has been to make interactions transparent by embedding the interface to intelligent systems in old and familiar objects, which

are therefore easy to use. A newer approach, however, has been to really dispense with objects altogether—to suffuse spaces with computation, therefore avoiding any point of interaction.

However, how many times have you seen hapless pedestrians stuck in front of an automatic "smart door"? They are unable to proceed because they don't know where the sensors, or the door's eyes, are located, and therefore they can't make the door open by making the right size of movements in the right quadrant. As Harry Potter says (Rowling 2000, p. 329), "Never trust anything that can think for itself, if you can't see where it keeps its brain."

Confusingly, projects involving "invisible computers" describe them as ways for people to interact with computation "as they interact with another person." As useful as is embedding computation in our environment, the notion must be tempered with knowledge of how humans actually do interact. We depend on forms of embodied interaction that offer us guidance in dealing with a complex world; interacting with invisibility does not fit one of the scripts. We need to locate intelligence, and this need poses problems for the invisible computer. The best example of located intelligence, of course, is the body. I'll talk about how the body ... *embodies* intelligence, both the usual knowledge about a particular domain and a less commonly discussed social interactional intelligence about conversational process, such as how to initiate, take turns, and interrupt in a conversation. In addition, I'll demonstrate how intelligent user interfaces can take advantage of embodied intelligence to facilitate human-machine interaction with a series of what I refer to as *embodied conversational agent* (ECA) systems.[1]

We need to locate intelligence, and this need poses problems for the invisible computer. The best example of located intelligence, of course, is the body.

An example of a person interacting with another person might serve to explain how humans actually do interact in their natural context and demonstrate some of the potential problems with interacting with invisibility. Figure 1 shows a young woman describing the layout of a house to a young man. Her eyes focus diagonally up and away as she plans her first utterance and then turn to her listener as she describes a complicated set up with her words and her hands. When the speaker says that the house is "surrounded by a porch all around," her hands demonstrate that the porch actually covers three sides of the house. The eye gaze toward the listener (depicted in the frozen frame in figure 1) elicits a feedback nod from him, during which the speaker is quiet (++ indicates silence). Once the speaker receives the listener's reaction, ensuring that speaker and listener share a common ground or understanding of what has already been described, she continues. She looks up as she plans her next utterance and repeats the gesture performance as she completes the description of the porch. The timing of the eye gaze, head movements, and hand gestures are tightly synchronized to speech, as marked by square brackets in the transcript. They are also tightly synchronized to the listener's behavior, as demonstrated by the feedback-eliciting gaze. The basic point is that people communicate with and to other people and not in a vacuum. Eyes gaze at other people and focus other people's attention on shared targets, hands gesture between people, faces express to other people. These behaviors are the external manifestations of social intelligence and trustworthiness (Cassell and Bickmore 2000) as well as a localization of the conversational processes of grounding information and a representation of information in their own right. Thus, if our goal is to reproduce how people communicate in natural contexts, we must also reproduce a way to localize the interaction and to represent the system's intelligence in space, to make the agency and intelligence of the participants visible by their actions and their reactions to communication.

Human Representation and Intelligence in Face-to-Face Conversation

The speaker in figure 1 knows something about the world that she is trying to convey to her listener, and she knows something about social conventions that is influencing how she goes about her task. We call these *propositional* and *interactional functions*, or skills. As described, both are carried by a number of behaviors in a number of different modalities: the voice, the hands, eye gaze, head movement.

We know that language is a representational medium (the *ur* representational medium), but are these other modalities anything other than fluff, pretty movements to occupy the body while the mouth is working? Eyes are not good representational tools (they can't describe), but they can certainly annotate (a discrete roll of the eyes while mentioning the election), focus the attention of one's interlocutor (as when one looks at one's own hands during a particularly complex gesture), and index appropriate social behavior (as earlier, when the speaker requests feedback by letting her gaze rest on her listener momentarily). Hands are excellent representational tools, better even than speech at representing the simultaneity of two events,

or the respective spatial locations of two objects ("so Lucy stood **there** and Betsy stood **there**"), and at disambiguating anaphoric reference ("and then **she** showed **her** how to move her feet"). In different contexts, gesture takes different forms: The more unfamiliar or surprising a speaker thinks a concept might be, the more representational the gesture accompanying mention of that concept (Cassell, Stone, et al. 2000). Thus, when talking to the human-computer interaction (HCI) community, a speaker might clasp his/her two hands together in front of him/her while saying the phrase *shared plans*. At a computational linguistics conference, a nod of the head in the direction of Barbara Grosz and Candy Sidner sitting in the audience would suffice. Both hands and head are skilled at taking up the slack of communication: a nod to acquiesce when one's mouth is too full to say "yes," a point toward one's full mouth to explain that one cannot speak. The body is the master of alternate and multiple representations, according to the needs and style of speaker and listener. Embodiment, therefore, would seem to fit the description of the ultimate interface, which "ultimately will include the ability to both retrieve and generate alternate representations of information according to the needs and personal styles of users" (Laurel 1990, p. 362).

Thus, these behaviors can both convey information and regulate communication in face-to-face conversation, but do they communicate; that is, does the listener pay any attention? In fact, yes, listeners depend on such embodied behaviors in face-to-face conversation. For example, they use the hand gesture they have seen in these situations to form a mental representation of the propositional content conveyed (Cassell, McNeill, et al. 1999), and they use the eye gaze to constrain when they make their own bids for the floor (Duncan 1974). We also know, however, that listeners are unable to remember what hand gestures they saw (Krauss, Morrel-Samuels, et al. 1991), and when they redescribe a monologue, they are likely to transpose the modality in which the information was conveyed. In addition, although teachers have been shown to use their pupils' gestures to judge the accuracy of the children's underlying understanding of mathematical concepts, they are unaware that they are so doing (Goldin-Meadow, Alibali, et al. 1993). Thus, just as with speech, the meanings underlying embodied interaction are extracted, but the behaviors themselves are not retained. However, when these embodied behaviors are omitted in face-to-face interaction between a user and an embodied system, users repeat themselves more and judge the system's use of language, and understanding of language, to be worse (Cassell and Thorisson 1999). In addition, when speech is ambiguous between humans (Thompson and Massaro 1986) or in a speech situation with some noise (Rogers 1978), listeners rely more on gestural cues (and, the higher the noise-to-signal ratio, the more facilitation by gesture). Thus, although the behaviors are not consciously retained, they are key to the interaction among humans, and between humans and machines. Note that the evidence presented thus far argues for different depictions on different modalities but one underlying modality-free common conceptual source that gives rise to the different instantiations, wherein each modality is called on to do what it does best. This semantic and pragmatic sharing of work recalls the interaction of words and graphics in an early kind of intelligent user interface—automatic generation of multimodal presentations (Feiner and McKeown 1991; Wahlster, Andre, et al. 1991—and recalls the separation of description and mechanism in Rosenschein and Kaebling's (1986) classic AI paper.

Figure 1. Describing a House.

The body is the master of alternate and multiple representations, according to the needs and style of speaker and listener. Embodiment, therefore, would seem to fit the description of the ultimate interface ....

From an ontological perspective, the importance of multiple representations in multiple modalities is not surprising. It has been argued that gestures are our first representational activity, arising from early sensorimotor schemata (Piaget 1952), and continue to replace unknown words in children's communication (Bates, Bretherton, et al. 1983); certainly, eye gaze and head movement regulate protoconversations between caregivers and infants before infants can even produce a semblance of language (Trevarthen 1986). Even in adults, nonverbal behaviors do not fade. About

The History of Embodied Interfaces: Automata

Wherefore are they endowed with organs so like to those of ourselves? Wherefore have they eyes, ears, nostrils, and a brain? It may be answered, that they may regulate the movements of the automata, by the different impressions which they receive from the exterior objects.

—*D'Alembert, 1717–1783*

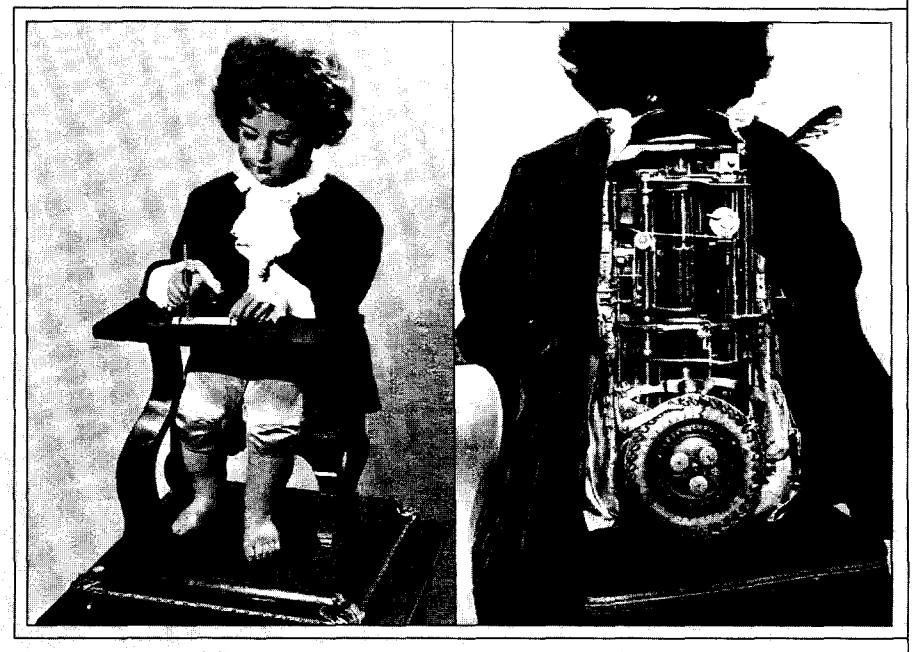

Figure A. Droz's Automaton, "The Writing Boy."

Attempts to model the body, and bodily interfaces, as well as attempts to make pretty bodies as entertainment systems, have been around for a very long time. In repudiation of Descartes's strict separation between the stuff that humans are made of and the thoughts they think, the organicist automaton makers of the eighteenth century asked whether one could design a machine that could talk, write, interact, play chess, and so forth, in the way people do. They intended to find out in this way what these activities consisted of when human beings perform them and how they differed, if at all, when machines perform them (Riskin 1999). For these reasons, designers in the organicist tradition tried to make their machines as much as possible like the organic subjects and processes they were imitating. Some organicist machines were strikingly lifelike representations of human processes and contributed significantly to knowledge about human functioning, for example, Droz's writing boy (figure A) whose pen moves across the page just as real writers' pens move.

Some organicist machines also acted as interfaces between humans and the world, for example, Von Kemplen's speaking machine (1791). Von Kemplen's machine (figure B) was the first that allowed users to produce not only some speech sounds but also whole words and short sentences. According to von Kemplen, it was possible to become a proficient user of the machine within three weeks and to then be able to produce strings of Latin, French, Italian, or German.

In contrast to these serious and scientific attempts to build embodiments and interfaces based on human functions were nineteenth-century automata that were meant to entertain, regardless of how humanlike

three-quarters of all clauses in narrative discourse are accompanied by gestures of one kind or another, regardless of cultural background (McNeill 1992).

The Conversational Model

Thus, humans engage in complex representational activity involving speech and hand gesture, and they regulate this activity through social conversational protocols that include speech and eye gaze and head movement and hand gesture. In this context, we can view the human in our earlier example as providing structure for her interlocutor that helps him navigate a complex description of the world. Her entire embodied performance provides cues to the shape of objects for which there is no adequate description in English, and to who has the floor. It also lets us know whether the two participants share common knowledge, and when the speaker's internal conceptual representation is in the process of being translated into words. Such a performance is helpful to the listener in understanding what is being

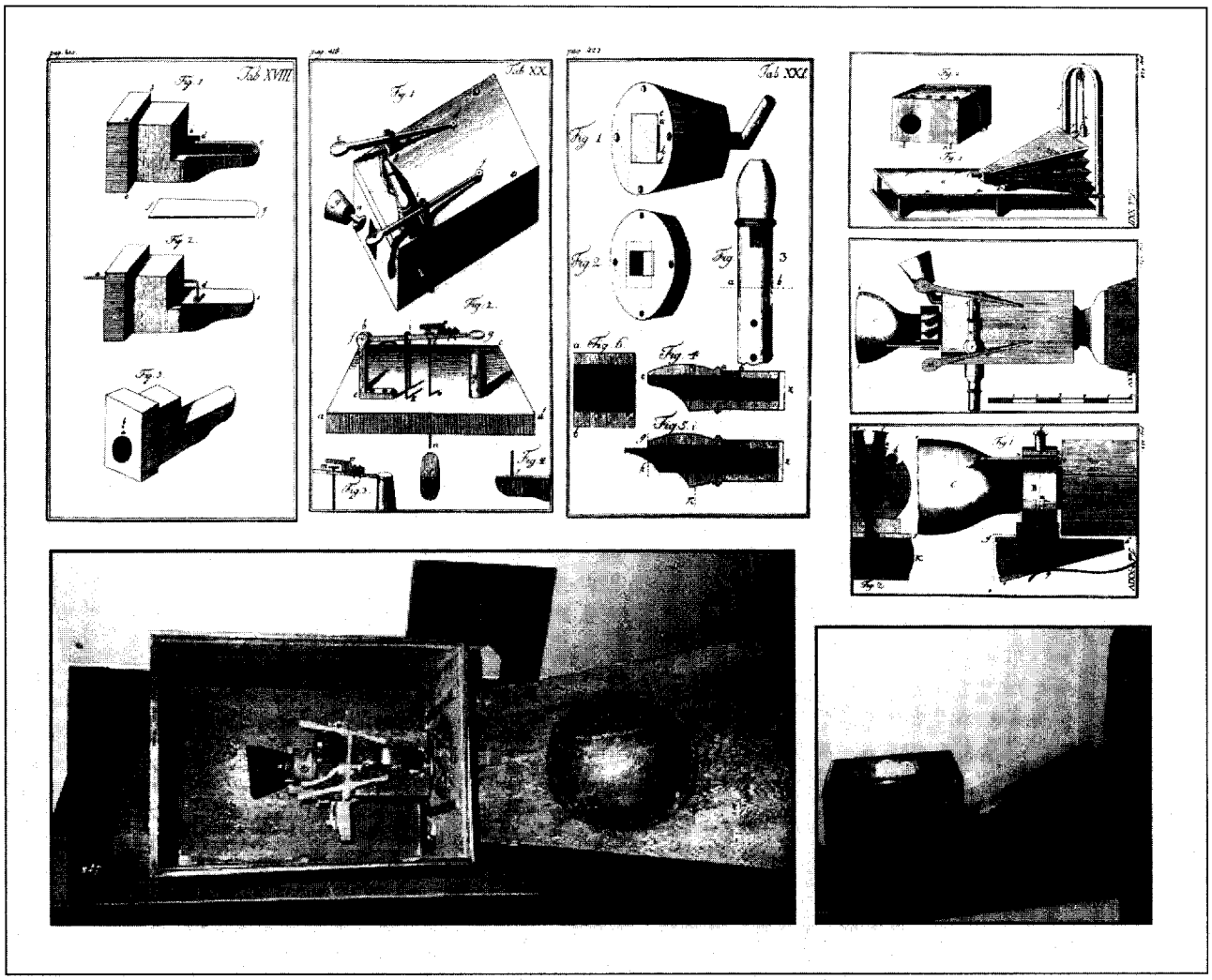

Figure B. Von Kemplen's Speaking Machine.

their actions might be. An example of such a pretty body as entertainment is the Pierrot automaton doll that writes—but simply by moving an inkless pen smoothly across a page—while it sighs deeply and progressively falls asleep by the lamplight.

Automaton makers were burned at the stake in the Middle Ages. Today in the interface community, we suspect some traditional human-computer interface researchers would be happy to do the same thing! As one of the most prominent critics has put it (Shneiderman 1998, p. 4),

> For those who built stone idols or voodoo dolls or the golem or Frankenstein, it's long been a dream.... But no mature technology resembles [animal] form. Automobiles don't run with legs, and planes don't flap their wings.... [Anthropomorphized agents] are things that think for people who don't.

Nevertheless, technology will most likely always be used to model humans to better understand how humans function and to leverage human understanding of the world by building that understanding into an interface that we learn to use from so early on in life.

said and integrating it into an ongoing discourse. It is also helpful in that it indicates that the speaker is a kind of representational device that the listener is familiar with. And it allows the listener to apply a theory of mind (Astington, Harris, et al. 1988) and, by doing so, to map the speaker's behaviors onto richer underlying representations, functions, and conventions—to attribute intelligence to the other.

In building an embodied conversational agent, I want both to help users steer their way through complex descriptions of the world and to prod them into automatically applying such a theory of mind as will allow them to not have

to spend their time constructing awkward new theories of the machine's intelligence on the fly. Thus, our model of conversational behavior must be able to predict exactly this kind of conversational behaviors and actions; that is, it must provide a way of realizing this set of conversational surface behaviors in a principled way.

This model, however rich, will not be able to predict all the behaviors displayed in human-human conversation, nor will it describe all the functions that give rise to these surface behaviors, nor would we wish it to. The model is not mimicking what people look like but adopting those aspects of the human interface that provide structure for our interpretations of meaning and for the process of interpreting meaning. Our goal is to target those behaviors that regulate and facilitate the process of interaction and represent information efficiently and effectively, all the while evoking a sense of another intelligence. Of course, however poor the model is, it will give rise to attributions that I have not planned: side-effects of cultural stereotypes of gender, race, and age that are evoked by the pitch of a voice, the tilt of a head, the form of a question (Reeves and Nass 1996). Steering our way through this Scylla and Charybidis of personification is helped by frequent evaluations of the system, in which users reveal the attributions—desired and not—that the system's behavior provokes.

These principles lead to a conversational model with several key properties: the system-internal representation of the world and information must be modality free but able to be conveyed by way of any one of several modalities; the functions of the system must be modality free but able to be realized in any one of a number of different surface behaviors in a number of different modalities; the representations of conversation cannot be all symbolic because cultural and social conventions might not be able to be captured in logical form; and co-occurrences of surface-level behaviors carry meaning over that carried by each of the constituent behaviors. In sum, I might describe such a model as a *multiplicity of representations*. We capture these properties and insights about human conversation in the FMBT (pronounced fembot) model.

F. Division between propositional and interactional functions: Contributions to the conversation are divided into propositional functions and interactional functions. The propositional function corresponds to the content of the conversation, including meaningful speech as well as hand gestures (gestures that indicate size in the utterance "it was this big" or that represent fingers walking in the utterance "it took me 20 minutes to get here"). The interactional function consists of cues that regulate the conversational process and includes a range of nonverbal behaviors (quick head nods to indicate that one is following, bringing one's hands to one's lap, and turning to the listener to indicate that one is giving up the turn) as well as regulatory speech ("huh?" "do go on"). In short, the interactional discourse functions are responsible for creating and maintaining an open channel of communication between the participants, and propositional functions shape the actual content. Both functions can be fulfilled with the use of a number of available communication modalities.

M. Modality: Both verbal and nonverbal modalities are responsible for carrying out the interactional and propositional functions. It is not the case that the body behaviors are redundant. The use of several different modalities of communication—such as hand gestures, facial displays, and eye gaze—is what allows us to pursue multiple goals in parallel, some of a propositional nature and some of an interactional nature. For example, a speaker can raise his/her pitch toward the end of the sentence while he/she raises the eyebrows to elicit feedback in the form of a head nod from the listener, all without interrupting the production of propositional content. It is important to realize that even though speech is prominent in conveying content in face-to-face conversation, spontaneous gesture is also integral to conveying propositional content. In fact, 50 percent of gestures add nonredundant information to the common ground of the conversation (Cassell, Stone, et al. 2000). For interactional communicative goals, the modality chosen might be more a function of what modality is free at a given point in the conversation; for example, is the head currently engaged in attending to the task, or is it free to give a feedback nod?

B. Behaviors are not functions: As can be seen from table 1, the same communicative function does not always map onto the same observed behavior. For example, the interactional function of giving feedback could either be realized as a head nod or a short "mhm." The converse is also true: The same behavior does not always serve the same function. For example, a head nod could be feedback or equally well a salutation or emphasis on a word. The particular set of surface behaviors exhibited can differ from person to person and from conversation to conversation (not to mention from culture to culture). Therefore, to successfully build a model of how conversation works, one cannot refer to these behaviors, or

Communicative Functions	Communicative Behavior
Initiation and termination:	
React to new person	Short glance at other
Break away from conversation	Glance around
Farewell	Look at other, head nod, wave
Turn-Taking:	
Give Turn	Look, raise eyebrows (followed by silence)
Want Turn	Raise hands into gesture space
Take Turn	Glance away,start talking
Feedback:	
Request Feedback	Look at other, raise eyebrows
Give Feedback	Look at other, nod head

Table 1. Some Examples of Conversational Functions and Their Behavior Realization
(from Cassell and Vilhjálmsson [1999]).

surface features, alone. Instead, the emphasis has to be on identifying the high-level structural elements or functions that make up a conversation. It is the understanding of these functions and how they work together to form a successful interaction that allows us to interpret the behaviors in context.

T. Time: Timing is a key property of human conversation, both within one person's conversational contributions and between participants. Within one person's contribution, the meaning of a nod is determined by where it occurs in an utterance, to the 200-millisecond scale. For example, consider the difference between "John [escaped]" (even though we thought it was impossible) and "[John] escaped" (but Bill did not). Between participants, a listener nod at precisely the moment a speaker requests feedback (usually by way of a rise in intonation and the flashing of eyebrows) is displaying understanding, but a delayed head nod might signify confusion. The rapidity with which behaviors such as head nods achieve their goals emphasizes the range of time scales involved in conversation. Although we have to be able to interpret full utterances to produce meaningful responses, we must also be sensitive to instantaneous feedback that can modify our interpretation and production as we go.

Although the FMBT model shares with Rodney Brooks and his colleagues a reliance on social interaction and a distinction between surface-level behaviors and underlying (deep structure) functions (Brooks, Brezeal, et al. 1998), in this model, these key properties do not displace the need for an explicit internal representation. Having a physical body, and experiencing the world directly through the influence of the world on that body, does not obviate the need for a model of the world. In Brooks's work, the autonomous systems he builds exploit features of the world and of humans to learn; the systems can go without a representation of their own so long as the world and humans manifest structure. In our work, humans exploit features of the interface to autonomous systems to achieve their goals; the interface must then present structure that the human can use. Intelligent user interfaces must provide representation. As Maybury and Wahlster (1998) remark, "A basic principle underlying multimedia systems is that the various constituents of a multimodal communication should be generated on the fly from a common representation of what is to be conveyed without using any preplanned text or images; that is, the principle is 'no generation without representation.'" Intelligent creatures can rely on the representations provided by others.

Rea: Implementing a FMBT-Embodied Conversational Agent

Thus far, I've talked about some of the properties of embodied human-human conversation that are essential for conveying information, regulating the course of the interaction, and giving one's interlocutor the sense that one is a familiar kind of representational creature. These key properties are captured in the FMBT intelligent interface model and distinguish the model from "intelligence without representa-

Figure 2. REA Welcoming a User to Her Virtual Realty Office.

tion" models of autonomous creatures. In this section, I give the details of how an embodied conversational agent can be implemented based on the model. To demonstrate, I turn to REA, an embodied conversational agent whose verbal and nonverbal behaviors are generated from underlying conversational functions and representations of the world and information. REA is the most extensive embodied conversational agent that my students and I have built on the basis of the FMBT model, which is why it is serving as an example. However, the architecture described here is independent of the REA implementation and has been used for a number of other embodied conversational agents (described more briefly in the last section).

First, REA has a humanlike body (shown in figure 2) and uses its body in humanlike ways during the conversation. That is, it uses eye gaze, body posture, hand gestures, and facial displays to contribute to the conversation and organize and regulate the conversation. It also understands (some aspects of the use of) these same modalities by its human interlocutor.

Second, the architecture allows for multiple threads of interaction to be handled, thus allowing REA to watch for feedback and turn requests. The human user can send such requests at any time through various modalities. The architecture is flexible enough to track these different threads of communication in ways appropriate to each thread. Because different threads have different response time requirements, the architecture allows different processes to concentrate on activities at different time scales.

Third, dealing with propositional information requires building a model of the user's needs and knowledge. Thus, the architecture includes both a static knowledge base that deals with the domain (here, real estate) and a dynamic discourse knowledge base (dealing with what has already been said). To generate propositional information, the system plans how to present multisentence multimodal output and manage the order of presentation of interdependent facts. To understand interactional information, on the other hand, the system builds 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).

Finally, the core modules of the system operate exclusively on functions (rather than sentences or behaviors, for example), while other modules at the edges of the system translate input into functions and functions into output. This division of labor also produces a symmetric architecture where the same functions and modalities are present in both input and output. Such models have been described for other conversational systems, for example, by Brennan and Hulteen (1995). I extend this previous research by developing a conversational model that relies on the functions of nonverbal behaviors, as well as speech, and that makes explicit the interactional and propositional contribution of these conversational behaviors.

Architecture

Figure 3 shows the modules of the REA architecture. Three main points translate the FMBT model for ECAs:

First, input is accepted from as many modalities as there are input devices. However the different modalities are integrated into a single conceptual representation that is passed from module to module.

Second, this conceptual representation frame has slots for interactional and propositional information so that the regulatory and content-oriented contribution of every conversational act can be maintained throughout the system.

Third, the categorization of behaviors in terms of their conversational functions is mirrored by the organization of the architecture that centralizes decisions made in terms of functions (the understanding, decision, and generation modules) and moves to the periphery decisions made in terms of behaviors (the input manager and action scheduler).

The *input manager* collects input from all modalities and decides whether the data require instant reaction or deliberate discourse processing. *Hard-wired reaction* handles rapid

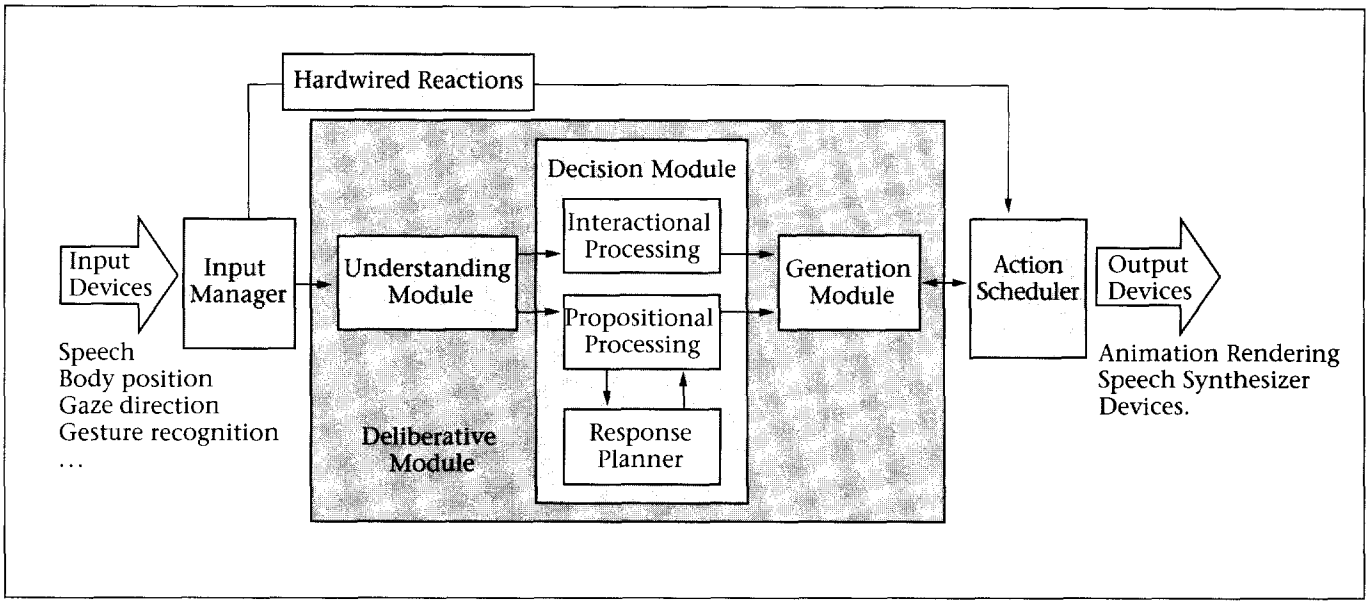

Figure 3. REA Architecture (codeveloped with the Fuji-Xerox Palo Alto Laboratory).

(under 200 milliseconds) reaction to stimuli, such as the appearance of the user. These stimuli can then directly affect the agent's behavior without much delay, which means that, for example, the agent's gaze can keep up with tracking the user's movement, without first processing the meaning of the user's appearance. The *deliberative discourse processing module* handles all input that requires a discourse model for proper interpretation, which includes many of the interactional behaviors as well as all propositional behaviors. Finally, the *action scheduler* is responsible for scheduling motor events to be sent to the animated figure representing the agent. A crucial function of the scheduler is to synchronize actions across modalities, so that, for example, gesture stroke and pitch peak in speech co-occur within milliseconds of each other. The modules communicate with each other using KQML, a speech-act–based interagent communication protocol, which serves to make the system modular and extensible.

Implementation

The system currently consists of a large back-projection screen on which REA is displayed and in front of which the user stands. Two cameras mounted on top of the projection screen track the user's head and hand positions in space. Users wear a microphone for capturing speech input. A single SGI OCTANE computer runs the conversation engine (originally written in C++ and CLIPS, currently moving to JAVA), and several other computers manage the speech recognition (until recently IBM VIA VOICE; currently moving to SUMMIT) and generation (previously Microsoft WHISPER; currently BT FESTIVAL), image processing (STIVE [Azarbayejani, Wren, et al. 1996] and VGUS [Campbell 2001], and graphics (written in OPENINVENTOR).

In the implementation of REA, we have attended to both propositional and interactional components of the conversational model, and all the modalities at REA's disposal (currently, speech with intonation, hand gesture, eye gaze, head movement, body posture) are available to express these functions. REA's current repertoire of interactional functions includes acknowledgment of a user's presence, feedback, and turn taking. Each of these functions is outlined as follows:

Acknowledgment: This is acknowledgment of a user's presence by turning to face the user.

Feedback: REA gives feedback in several modalities: It might nod its head or emit a paraverbal (for example, "mmhmm") or a short statement such as "okay" in response to short pauses in the user's speech; it raises its eyebrows to indicate partial understanding of a phrase or sentence.

Turn taking: REA tracks who has the speaking turn and only speaks when it holds the turn. Currently, REA always allows verbal interruption and yields the turn as soon as the user begins to speak. If the user gestures, it will interpret this gesture as an expression of a desire to speak and, therefore, will halt its remarks at the nearest sentence boundary. Finally, at the end of its speaking turn, it turns to face the user.

These conversational functions are realized

State	Output Function	Behaviors
User Present	Open interaction	Look at user. Smile. Toss head.
	Attend	Face user.
	End of interaction	Turn away.
	Greet	Wave. Say "hello.".
REA Speaking	Give turn	Relax hands. Look at user. Raise eyebrows.
	Signoff	Wave. Say "bye."
User Speaking	Give feedback	Nod head, paraverbal ("hmm").
	Want turn.	Look at user. Raise hands.
	Take turn.	Look at user. Raise hands to begin gesturing. Speak.

Table 2. Output Functions.

In terms of the propositional component, REA's speech and gesture output are generated in real time, and words and gesture are treated on a par, so that a gesture can be just as likely to be chosen to convey REA's meaning as a word.

as conversational behaviors. For turn taking, for example, the specifics are as follows: REA generates speech, gesture, and facial expressions based on the current conversational state and the conversational function it is trying to convey. For example, when the user first approaches REA (user present state), it signals its openness to engage in conversation by looking at the user, smiling, and tossing its head. When conversational turn taking begins, it orients its body to face the user at a 45-degree angle. When the user is speaking, and REA wants the turn, it looks at the user. When REA is finished speaking and ready to give the turn back to the user, it looks at the user, drops its hands out of gesture space, and raises its eyebrows in expectation. Table 2 summarizes REA's current interactional output behaviors.

In terms of the propositional component, REA's speech and gesture output are generated in real time, and words and gesture are treated on a par, so that a gesture can be just as likely to be chosen to convey REA's meaning as a word. The descriptions of the houses that it shows, along with the gestures that it uses to describe these houses, is generated using the SPUD natural language–generation engine, modified to also generate natural gesture (Cassell, Stone, et al. 2000). New propositional information is conveyed using iconic gestures (for concepts with concrete existence), metaphoric gestures (for concepts that do not have concrete existence and, thus, must make use of spatial metaphors for depiction), or deictic gestures (for indicating or emphasizing an object in REA's virtual world, such as features of homes it is showing to the user). These gestures are either wholly redundant with, or complementary to, the speech channel based on semantic and pragmatic constraints. Beats are used to indicate points of emphasis in the speech channel without conveying additional meaning.

When REA produces an utterance, then it first determines several pieces of pragmatic and semantic information, including *semantics*, which is a speech act description of REA's communicative intent (for example, **offer** the user a particular property, **describe** a room); *information structure,* which tells which entities are new versus previously mentioned; *focus,* which tells which entity (if any) is currently in focus; and *mutually observable information,* which tells which entities in the virtual world are visible to both REA and the user.

This information is then passed to the SPUD unified text- and gesture-generation module that generates REA's natural language responses. This module distributes the information to be conveyed to the user across the voice and gesture channels based on the semantic and pragmatic criteria described earlier and is timed so that gestures coincide with the new material in the utterance. If a new entity is in focus and it is mutually observable, then a deictic is used. Otherwise, REA determines if the semantic content can be mapped into an iconic or metaphoric gesture (using heuristics derived from studies of the gestures humans produce in describing real estate [Yan 2000] to determine

whether the gestures should be complementary or redundant). For example, REA can make a walking gesture (extending its index and second finger with the tips downward, as if they are legs, and wiggling the fingers back and forth) as it says, "The house is five minutes from MIT." In this case, the gesture carries complementary information—that the house is five minutes on foot rather than five minutes by car. REA can also make a sweeping "sun-rising" gesture with both arms above her head, as she says, "the living room is really luminous." In this case, the gesture is redundant to the notion of sunniness conveyed by speech.

REA is also able to detect these same classes of gestures made by the user and combine this information with speech input to interpret propositional content and make decisions about appropriate responses. Here, once again, our reliance on conversational function, and distinction between function and surface behavior, allows us to bypass questions of gestural form (a difficult vision problem indeed) and concentrate on how the user is using a given gesture in the current conversational context. The gesture-classification module uses a set of hidden Markov model (HMM) recognizers to classify gestures into the categories of rest (no gesture), beat, deictic, butterworth (searching for a word), or illustrative (iconic or metaphoric). The HMMs that classify into these categories were trained in an offline process from a set of 670 gestures obtained by tracking naive subjects with STIVE as they engaged in real-estate–oriented conversations and then hand segmenting and classifying the subjects' conversational gestures (Campbell 2001).

To date, REA is only capable of integrating the beat and deictic categories into the decision module. The beat category is used to interpret user emphasis with respect to the speech channel, allowing REA to know which part of a user utterance to respond to (if the user says "I like the blue tiles" with a beat gesture on the word *blue,* then REA responds by saying, "Blue is my favorite color." If the user says, "I like the blue tiles" with a beat gesture on the word *tiles,* then REA responds by saying, "I love tiles."). REA can also pick up when users point to objects in REA's world and can associate the deictic gesture and the graphical object pointed at with the word that co-occurs with the deictic, thus enabling REA to resolve a wider range of referring expressions ("that house," "the wall," and so on). These functions are illustrated in the example interaction sidebar.

By modeling behavioral categories as discourse functions, we have developed a natural and principled way of combining multiple modalities in both input and output. Thus, when REA decides to give feedback, for example, it can choose any of several modalities based on what is appropriate and available at the moment. When it conveys its knowledge of the

An Example Interaction

The following records an actual interaction between a user and REA:
Mike approaches the projection screen. REA is currently turned side on and is gazing idly about. As Mike moves within range of the two cameras mounted above the screen, REA turns to face him and says:
REA> Hello. How can I help you?
MIKE> I'm looking to buy a place near MIT.
REA nods, indicating that it is following.
REA> I have a house to show you.
A picture of a house appears on a screen behind REA, who blinks and looks at the house and then at Mike.
REA> It's in Somerville.
MIKE> Tell me about it.
REA looks up and away while it plans what to say.
REA> It's big.
REA makes an expansive gesture with her hands.
Mike brings his hands up as if he is about to speak and so REA does not continue, instead waiting for him to speak.
MIKE> Tell me more about it
REA> Sure thing. It has a nice garden.
REA sketches a curved gesture with her hands indicating that the garden extends along two sides of the house.
MIKE> How far is it?
REA> It is five minutes to the Porter Square T station.
REA makes it clear that it is five minutes on foot from the T station by making a walking gesture with its fingers.
MIKE> How big is the house?
REA> It has four bedrooms, three bathrooms....
Mike interrupts REA, who stops speaking immediately.
MIKE> Wait. Tell me. Where is the master bedroom?
REA> I'm sorry, I didn't catch that. What did you ask me?
MIKE> Where is the master bedroom?
REA> It's upstairs.
REA points up.
MIKE> Where is the master bathroom?
REA> It's next to the bedroom.
REA brings her hands together to indicate the relationship between the bedroom and the bathroom.
The house tour continues....

Figure 4. ANANOVA.

world, all its behaviors are marshaled toward giving a well-rounded description of what it knows.

Related Work

Humanlike embodied interfaces have become popular as the front end to commercial systems; so, it doesn't seem that what is described here is altogether novel. However, although these interfaces look like bodies, few of them display behaviors or manifest the types of functions of bodies in conversation that I've argued for. Surprisingly, in fact, not much has changed since a CHI panel about anthropomorphism in 1992 was advertised in the following way: "Recently there has been a discernible increase in the gratuitous use of the human figure with poorly lipsynched talking heads or systems that fool the user into thinking that the system is intelligent" (Don, Brennan, et al. 1992). Well, actually, lipsynching has improved.

Well-designed interfaces have affordances or visual clues to their operation. Bodies provide strong visual clues about the protocols that they engage in, and these protocols must be integrated into the very heart of a system and must give rise to appropriate surface-level behaviors in the interface for the embodied interface to be successful. In contradistinction to this methodology, many of the humanlike interfaces on the market simply consist of an animated character slapped onto a system, capable of portraying a series of affective or "communicative" poses, without much attention paid to how humans actually convey their knowledge of the world and of human interaction to their interlocutors. Two examples of less than perfect embodied interfaces that come to mind are the Microsoft OFFICE ASSISTANT (the dreaded PAPER CLIP) and ANANOVA (figure 4). The PAPER CLIP (or the more anthropomorphic version EINSTEIN) interrupts in an impolite and socially inappropriate manner and, when not actually typing, manifests its profound boredom in the user's work by engaging in conversationally irrelevant behaviors (as if one's interlocutor checked his/her watch while one was speaking and then snapped to attention when it was his/her turn to talk). Interestingly, interruption itself is not the problem; participants in conversations interrupt one another all the time. However, interruption must be motivated by the demands of the conversation—requests for further information, excitement at what is being said—and must follow the protocol of conversation (for example, raise the hands into gesture space, clear the throat, extend the feedback noise for longer than usual as ways of requesting the floor). ANANOVA is advertised as a way to personalize web users' interaction with information but is not, in fact, capable of interaction (or, currently, personalization). This not uncommon confusion of personalization, and graphical representation of a person leads to putting a body on the interface for looks and personality as opposed to for the functions of the body. ANANOVA sways slightly and winks or sneers occasionally but does not pause or request feedback, check its viewer's response to what it is saying, or in any other way attend to its viewers. It also does not use any modalities other than voice to convey content. The pages advertised as showing "technical drawings" tell us, "Then one of [my creators] had the bright idea [to] unleash my full potential by giving me a human face and full-rounded personality so that I could better interact with people as technology develops." As far as one can tell, ANANOVA's behaviors are hand scripted as annotations to text (for example, it's hard to imagine the set of underlying rules for conversation and representation for information that would make her lips curl slightly in a sneer when it says, "I've been locked in a room for 12 months with only geeks and techies for company"). Even when the day comes where it is able to deliver all and only the news that a particular web viewer requests, its interaction with this web user will not resemble conversational interaction, and its embodiment will not make it appear any more intelligent an interface.

Such systems represent an enormous missed opportunity. As I've argued, used appropriately, an embodied interface can provide not only something pretty and entertaining to look at but can also enable the use of certain communication protocols in face-to-face conversation that facilitate user interaction and provide for a more rich and robust channel of communication than is afforded by any other mediated