ELSEVIER

Knowledge-Based Systems 14 (2001) 55–64

Knowledge-Based SYSTEMS

www.elsevier.com/locate/knosys

More than just a pretty face: conversational protocols and the affordances of embodiment

J. Cassell*, T. Bickmore, L. Campbell, H. Vilhjálmsson, H. Yan

Gesture and Narrative Language Group, MIT Media Laboratory, E15-315, 20 Ames St, Cambridge, MA, USA

Abstract

Prior research into embodied interface agents has found that users like them and find them engaging. However, results on the effectiveness of these interfaces for task completion have been mixed. In this paper, we argue that embodiment can serve an even stronger function if system designers use actual human conversational protocols in the design of the interface. Communicative behaviors such as salutations and farewells, conversational turn-taking with interruptions, and describing objects using hand gestures are examples of protocols that all native speakers of a language already know how to perform and can thus be leveraged in an intelligent interface. We discuss how these protocols are integrated into Rea, an embodied, multi-modal interface agent who acts as a real-estate salesperson, and we show why embodiment is required for their successful implementation. © 2001 Elsevier Science B.V. All rights reserved.

Keywords: Rea; Embodied interface agent; Communicative behavior; Embodied conversational agent

1. Introduction

There is a qualitative difference between face-to-face conversation and other forms of human–human communication [1]. Businesspeople and academics routinely travel long distances to conduct certain face-to-face interactions when electronic forms of communication would seemingly work just as well. When someone has something really important to say, they say it in person.

The qualitative difference in these situations is not just that we enjoy looking at humans more than at computer screens but also that the human body enables the use of certain communication protocols in face-to-face conversation which provide for a more rich and robust channel of communication than is afforded by any other medium available today. The use of gaze, gesture, intonation, and body posture play an essential role in the proper execution of many conversational functions — such as conversation initiation and termination, turn-taking, interruption handling, feedback and error correction — and these kinds of behaviors enable the exchange of multiple levels of information in real time. People are extremely adept at extracting meaning from subtle variations in the performance of these behaviors; for example slight variations in pause length, feedback nod timing or gaze behavior can significantly alter the interpretation of an utterance (consider "you did a great job" vs. "you did a … great job").

Of particular interest to interface designers is that these communication protocols come for "free" in that users do not need to be trained in their use; all native speakers of a given language have these skills and use them daily. An embodied interface agent which exploits these protocols has the potential to provide a higher bandwidth of communication than would otherwise be possible.

Of course, depictions of human bodies are also more decorative than menus on a screen and, like any new interface design, they are also currently quite in vogue and therefore attractive to many users. Unfortunately, many embodied interface agents developed to date don't go further than their ornamental or novelty value. Aside from the use of pointing gestures and two or three facial expressions, an extensive wardrobe and a coyly cocked head, many animated interface agents provide little more than something amusing to look at while the same old system handles the mechanics of the interaction. It is no wonder that these systems have been found to be likable and engaging, but to provide little improvement in task performance over text or speech-only interfaces.

In this paper, we first review the embodied interface agents developed to date and summarize the results of evaluations performed on them. We then discuss several human communication protocols along with their interface utility and requirements for embodiment. Finally, we

* Corresponding author. Tel.: +1-617-253-4899.
E-mail addresses: justine@media.mit.edu (J. Cassell), bickmore@media.mit.edu (T. Bickmore), elwin@media.mit.edu (L. Campbell), hannes@media.mit.edu (H. Vilhjálmsson), yanhao@media.mit.edu (H. Yan).

PII: S0950-7051(00)00102-7

present Rea, an embodied interface agent which implements these protocols and describe our ongoing research program to develop embodied interface agents that leverage knowledge of human communication skills.

2. Related work

Other researchers have built embodied interface agents, with varying degrees of conversational ability. Olga is an embodied humanoid agent that allows the user to employ speech, keyboard and mouse commands to engage in a conversation about microwave ovens [2]. Olga has a distributed client–server architecture with separate modules for language processing, interaction management, direct manipulation interface output animation, all communicating through a central server. Olga is event driven, and so only responds to user input and is unable to initiate output on its own. In addition, Olga does not support non-speech audio or computer vision as input modalities.

Olga uses a linear architecture in which data flows from user input to agent output, passing through all the internal modules in between. Takeuchi and Nagao [3] suggest a different approach. Their conversational agent is based on the subsumption architecture by Rodney Brooks [4]. In this case the agent is based on a horizontal decomposition of task-achieving behavior modules. The modules each compete with one another to see which behavior is active at a particular moment. Thus there is no global conversational state or model and the conversational interaction arises from the interplay between the different behavioral layers. Their agent responds to speech and gaze information, but coordination of the input analysis and output generation is also an emergent behavior, so precise control is impossible. The end result is that user input and agent output are decomposed according to task behaviors rather than conversational function.

Lester et al. [5] do generate verbal and non-verbal behavior, producing deictic gestures and choosing referring expressions as a function of the potential ambiguity of objects referred to, and the proximity of those objects to the animated agent. This system is based on an understanding of how reference is achieved to objects in the physical space around an animated agent, and the utility of deictic gestures in reducing potential ambiguity of reference. However, the generation of gestures and the choice of referring expressions (from a library of voice clips) are accomplished in two entirely independent (additive) processes, without a description of the interaction between the two modalities. Likewise, Rickel and Johnson [6] have their pedagogical agent move to objects in the virtual world that it inhabits, and then generate a deictic gesture at the beginning of the verbal explanation that the agent provides about that object. Andre and Rist (this volume) generate verbal and non-verbal information that is presented by two embodied agents speaking to one another. Their system generates different displays based on the personality and attitude of each agent. These last three systems are closest to our own research in this area. In these systems, however, the association between verbal and non-verbal behaviors is additive — that is, the information conveyed by hand gestures, for example, is always redundant with the information conveyed by speech. The affordances of the body are not exploited for the kinds of tasks that it performs better than speech.

"Animated Conversation" [7] was a system that automatically generated context-appropriate gestures, facial movements and intonational patterns. In this case the domain was conversation between two artificial agents and the emphasis was on the production of non-verbal propositional behaviors that emphasized and reinforced the content of speech. However, the system was not designed to interact with a user, and did not run in real time.

The work of Thorisson provides a good first example of how an embodied interface agent inspired by studies of human psychosocial competencies might be developed [8]. The agent, Gandalf, recognized and displayed interactional information such as gaze, simple gesture and canned speech events. In this way he was able to perceive and generate turn-taking and back channel behaviors that lead to a very natural conversational interaction. However, Gandalf had limited ability to recognize and generate propositional information, and was also limited in his ability to provide correct intonation for speech emphasis on speech output, or co-occurring gestures with speech.

The conversational character system developed by Prevost et al. [9], uses the same architecture as the one presented in this paper (it was co-developed by our two research groups), but their application domain and many implementation details are different. In their system a conversational character assists a user with a complex A/V system by controlling equipment, answering questions or giving tutorials. To date, the conversational behaviors of their agents are limited to greeting and farewell rituals, gaze, pointing gestures and body positioning.

In another vein entirely, research on generating words and graphics in multi-modal presentations [10–12] (Kerpedjiev and Roth, this volume) examines the different affordances of language and pictures, and reminds us that in writing as in speaking communicative goals may be differentially mapped onto different modalities.

2.1. User studies on embodied interface agents

Koda and Maes [13], and Takeuchi and Naito [14], studied user responses to interfaces with static or animated faces, and found that users rated them to be more engaging and entertaining than functionally equivalent interfaces without a face. Kiesler and Sproull found that users were more likely to be cooperative with an interface agent when it had a human face (vs. a dog image or cartoon) [15].

Andre et al. found that users rated their animated

Table 1
The FEMBOT model of embodied conversation

FEMBOT model
● F: Propositional and interactional functions
● M: Multi-modal (speech, gesture, eye gaze …)
● B: Separation of function and behavior
— Simplifies implementation
— Allows modularity with respect to personality and culture
● T: Real-time
— Attention paid to overall responsiveness
— Tight temporal synchronization in input and output

presentation agent ("PPP Persona") as more entertaining and helpful than an equivalent interface without the agent [16]. However, there was no difference in actual performance (comprehension and recall of presented material) in interfaces with the agent vs. interfaces without it.

In a user study of the Gandalf system mentioned above [17], users rated the smoothness of the interaction and the agent's language skills significantly higher under test conditions in which Gandalf utilized limited conversational behavior (gaze, turn-taking and limited gesture) than when these behaviors were disabled.

Most of these evaluations have tried to address whether embodiment of a system is useful at all, usually by keeping the interaction the same, and then including or not including an animated figure. The studies, then, are not testing how particular uses of embodiment may improve task or learning performance. Therefore, although the previous studies inspire us by showing that the mere presence of a character wins us points, we now need to focus on the contribution of embodiment in fully functional conversational interfaces, and in order to do that, we need to start with a better understanding of what embodiment contributes to human–human interaction.

Table 2
Some examples of conversational functions and their behavior realization

Communicative functions	Communicative behavior
Initiation and termination	
Reacting	Short glance
Inviting Contact	Sustained glance, Smile
Distance salutation	Looking, Head toss/Nod, Raise eyebrows, Wave, Smile
Close salutation	Looking, Head nod, Embrace or handshake, Smile
Break away	Glance around
Farewell	Looking, Head nod, Wave
Turn-taking	
Give turn	Looking, Raise eyebrows (followed by silence)
Wanting turn	Raise hands into gesture space
Take turn	Glance away, Start talking
Feedback	
Request feedback	Looking, Raise eyebrows
Give feedback	Looking, Head nod

3. Human communication protocols requiring embodiment

Providing the interface with a body allows the system to engage in a wide range of multi-modal behaviors that, when executed in tight temporal synchronization with language, carry out a communicative function. It is important to understand that particular behaviors, such as the raising of the eyebrows, can be employed in a variety of circumstances to realize different communicative functions, and that the same communicative function may be realized through different sets of behaviors. It is therefore important for any system dealing with conversational modeling to handle function separately from surface-form or run the risk of being inflexible and insensitive to the natural phases of the conversation. This distinction between form to function relies on a fundamental division of conversational goals: contributions to a conversation can be propositional and interactional. Propositional information corresponds to the content of the conversation. This includes meaningful speech as well as hand gestures (gestures that indicate the size in the sentence "it was this big"). Interactional information consists of the cues that regulate conversational process and includes a range of non-verbal behaviors (quick head nods to indicate that one is following) as well as regulatory speech ("huh?", "uh-huh"). This theoretical stance allows us to examine the role of embodiment not just in task — but also process-related behaviors. From this standpoint, we note that most previous embodied interface agents do not deal with interactional and propositional information in an integrated manner, which prevents them from fully exploiting the affordances of the body. We capture these insights in a model of embodied conversation called the FEMBOT model, as shown in Table 1.

Below we briefly describe some of the fundamental communication protocols and their functional elements along with examples of non-verbal behavior that contribute to their successful implementation. Table 2 shows examples of mappings from communicative function to particular behaviors and is based on previous research on typical North American non-verbal displays, mainly [18] and [19].

3.1. Conversation initiation and termination

Humans partake in an elaborate ritual when engaging and disengaging in conversations [19]. For example, people will show their readiness to engage in a conversation by turning towards their potential interlocutors, gazing at them and then exchanging signs of mutual recognition typically involving a smile, eyebrow movement and tossing the head or waving of the arm. Following this initial synchronization stage, or distance salutation, the two people approach one another, sealing their commitment to the conversation through a close salutation such as a handshake accompanied by a ritualistic verbal exchange. The greeting phase ends when the two participants re-orient their bodies, moving

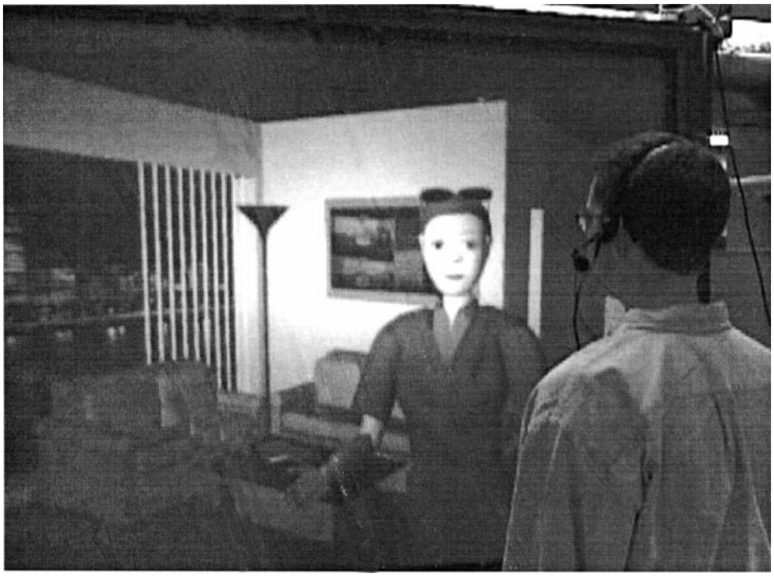

Fig. 1. User interacting with Rea.

away from a face-on orientation to stand at an angle. Terminating a conversation similarly moves through stages, starting with non-verbal cues, such as orientation shifts or glances away and cumulating in the verbal exchange of farewells and the breaking of mutual gaze.

3.2. Conversational turn-taking and interruption

Interlocutors do not normally talk at the same time, thus imposing a turn-taking sequence on the conversation. The protocols involved in floor management — determining whose turn it is and when the turn should be given to the listener — involve many factors including gaze and intonation [20]. In addition, listeners can interrupt a speaker not only with voice, but also by gesturing to indicate that they want the turn. Floor management is not simply a question of *recognizing* the behavior of one's conversational partner; it is a truly *responsive* or co-constructed activity [21].

3.3. Content elaboration and emphasis

Gestures can convey information about the content of the conversation in ways that the hands are uniquely suited to meet. For example, the two hands can better indicate simultaneity and spatial relationships than the voice or other channels. Probably the most commonly thought of use of the body in conversation is the pointing (deictic) gesture, possibly accounting for the fact that it is also the most commonly implemented for the bodies of animated interface agents. In fact, however, most conversations don't involve many deictic gestures [22] unless the interlocutors are discussing a shared task that is currently present. Other conversational gestures also convey semantic and pragmatic information. Beat gestures are small, rhythmic baton like movements of the hands that do not change in form with the content of the accompanying speech. They serve a pragmatic function, rather like intonational prominence, conveying information about what is "new" in the speaker's discourse. Iconic and metaphoric gestures convey some features of the action or event being described [22]. They can be redundant or complementary relative to the speech channel, and thus can convey additional information or provide robustness or emphasis with respect to what is being said. Whereas iconics convey information about spatial relationships or concepts, metaphorics represent concepts that have no physical form, such as a sweeping gesture accompanying "the property title is free and clear."

3.4. Feedback and error correction

During conversation, speakers can non-verbally request feedback from listeners through gaze and raised eyebrows and listeners can provide feedback through head nods and paraverbals ("uh-huh", "mmm", etc.) if the speaker is understood, or a confused facial expression or lack of positive feedback if not. Listeners can also ask clarifying questions if they did not hear or understand something the speaker said.

4. Rea: an embodied conversational agent

The Rea project at the MIT Media Lab [23,24] has as its goal the construction of an embodied, multi-modal real-time conversational interface agent. Rea implements the conversational protocols described above, on the basis of the FEMBOT model, in order to make interactions as natural as face-to-face conversation with another person. In the current task domain, Rea acts as a real estate salesperson, answering user questions about properties in her database and showing users around virtual houses.

Rea has a fully articulated graphical body, can sense the

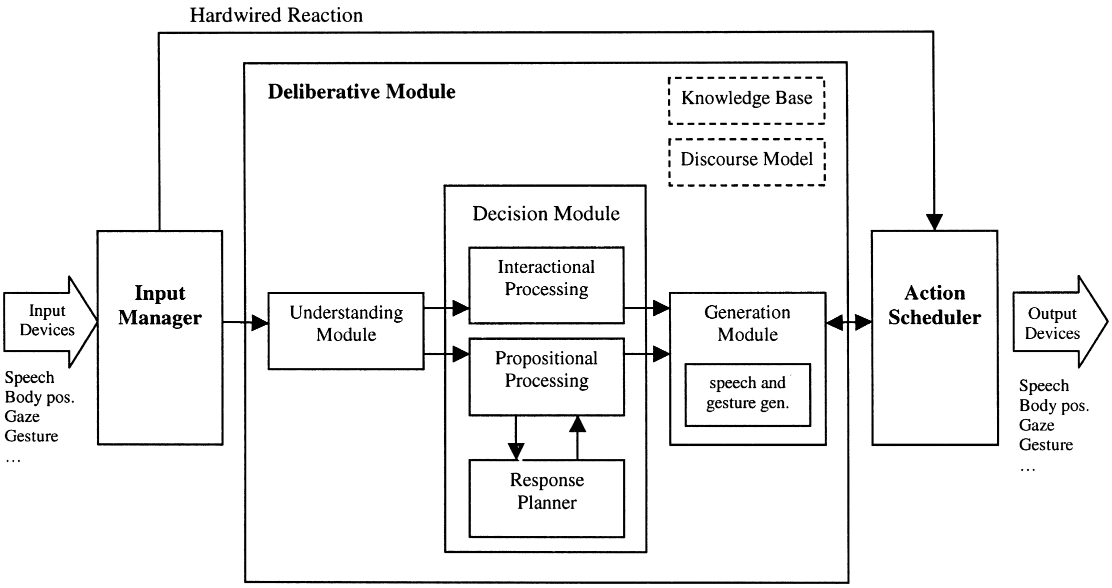

Fig. 2. Rea's software architecture.

user passively through cameras and audio input, and is capable of speech with intonation, facial display, head and eye movement, and gestural output. The system currently consists of a large projection screen on which Rea is displayed and which the user stands in front of. 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 graphics and conversation engine of Rea, while several other computers manage the speech recognition and generation and image processing. Fig. 1 shows Rea in action.

Rea is able to conduct a conversation describing the features of the task domain while also responding to the users' verbal and non-verbal input. When the user makes cues typically associated with turn taking behavior such as gesturing, Rea allows herself to be interrupted, and then takes the turn again when she is able. She is able to initiate conversational error correction when she misunderstands what the user says, and can generate combined voice, facial expression and gestural output. Rea's responses are generated by an incremental natural language generation engine based on [25] that has been extended to synthesize redundant and complementary gestures synchronized with speech output [26]. A simple discourse model is used for determining which speech acts users are engaging in, and resolving and generating anaphoric references.

4.1. Architecture

Fig. 2 shows the modules of the Rea architecture that is designed to meet the requirements of real-time face-to-face conversation [23]. In this design, input is accepted from as many modalities as there are input devices. However the different modalities are integrated into a single semantic representation that is passed from module to module. This representation is a KQML frame [27] that 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.

The categorization of behaviors in terms of their conversational functions is mirrored by the organization of the architecture which centralizes decisions made in terms of functions (in the Deliberative Module), and moves to the periphery decisions made in terms of behaviors (the Input Manager (IM) and Action Scheduler (AS)).

In addition the IM and AS can communicate through a hardwired reaction connection, to respond immediately (under 200 ms) to user input or system commands. Tracking users with gaze shifts as they move is an example of a reactive behavior. The other modules are more "deliberative" in nature and perform non-trivial inferencing actions that can take multiple real-time cycles to complete. Rea is implemented in C++ and CLIPS, a rule-based expert system language [28].

4.2. Overview of implemented communication protocols

Rea implements the human communication protocols previously described, as follows.

4.2.1. Conversation initiation and termination

Rea acknowledges the user's presence through posture, by turning to face the user, as detected by the vision system. She also exchanges greetings and farewells with the user using verbal and non-verbal (gestural) output, in response to the user's verbal greeting and farewell. Rea also recognizes when the user turns away during conversation (based on vision input) and suspends speech input processing until the user turns to face her again.

4.2.2. Conversational turn-taking and interruption

Rea tracks who has the speaking turn (using a conversational state model), and only speaks when she holds the turn. Currently Rea always allows verbal interruption based on audio threshold detection and yields the turn as soon as the user begins to speak. If the user gestures (as detected by the vision system) she will interpret this as expression of a desire to speak, and halt her remarks at the nearest sentence boundary. She exhibits the "look away" behavior while she is planning her response (which serves to hold the turn until she is ready to speak), and at the end of her speaking turn she turns to face the user to indicate a turn transition point.

4.2.3. Content elaboration and emphasis

Rea currently can generate a wide range of gestures to both convey propositional information and to emphasize information in her speech. New propositional information is conveyed using *iconic* gestures (for concepts with concrete existence), *metaphoric* gestures (for concepts which 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 she is showing to the user). These gestures may be wholly redundant with or complementary to the speech channel. *Beats* are used to indicate points of emphasis in the speech channel without conveying additional meaning.

When Rea decides to produce an utterance, she first determines several pieces of pragmatic and semantic information required to generate speech and gesture, including:

- semantics — speech act description of Rea's communicative intent (e.g. OFFER the user a particular property, DESCRIBE a room, etc.);
- information structure — which entities are new (rheme) vs. previously mentioned (theme);
- focus — which entity (if any) is currently in focus; and
- mutually observable — which entities in the virtual world are visible to both Rea and the user.

This information is then passed to a unified text generation module (GM) [25,26] which generates Rea's natural language responses together with accompanying conversational gestures. 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 above. Gestures are placed to coincide with rhematic 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 [29]) to determine whether the gestures should be complementary or redundant. For example, Rea may make a *walking gesture* (extending her index and second finger with the tips downward, as if they are legs, and wiggling the fingers back and forth) as she says "It's 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. Or, Rea may 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. If none of these cases can be realized, then the module introduces a beat gesture at the appropriate point.

The text GM outputs an utterance annotated with gestures specified compositionally as a function of hand starting and ending positions, trajectory, hand shape and envelope size. A scheduling module then estimates phoneme timings, maps the gesture specifications to animation primitives, adds any necessary preparatory, retraction, and co-articulation primitives, and prepares a complete execution plan for the utterance before it is passed to an animation module for performance.

Rea is also able to detect certain classes of gestures made by the user and combine this information with speech input to interpret messages and make decisions about appropriate responses. The gesture classification module is based on input from STIVE [30], the vision system, and a prototype of this GESTIRP [31] module is running in the current version of REA, and classifying gestures as they occur. It obtains 3D coordinates of the hands and head as they move over time, transforms them into velocities in a user body-centered coordinate system, and classifies sequences of velocity measurements using a set of Hidden Markov Model (HMM) [32] recognizers.

The HMMs classify the gestures into one of the following seven categories: rest (no gesture), beat, preparation, retraction, 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 [31].

So far, REA only uses the beat, preparation, and retraction categories in the conversational planning process. The preparation and retraction categories prevent the movement from being misunderstood as some other gesture, and the beat category is used to interpret user emphasis with respect to the speech channel (see extended example below). However, we are currently implementing the ability to recognize users' deictic gestures when they point to objects in REA's world, and to associate both 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", etc.).

```
(tell :sender UM :recipient DM :content
  (commact :sender USER :recipient REA
    :input [(speaking :state TRUE)
            (gesturing :state TRUE) ]
    :prop NONE
    :intr [ (takingturn) ]
  )
)
```

Fig. 3. A sample performative.

4.2.4. Feedback and error correction

Rea provides non-verbal feedback during the user's turn by nodding her head at the end of user utterances (as detected by the audio threshold device) in which the user keeps the turn. If Rea does not fully understand the user's input (typically due to errors reported by the speech recognition system) she attempts repair by asking a clarifying question.

4.3. Sample interactions in detail

In order to understand better how Rea processes user input, both propositional and interactional, and produces appropriate output behavior, it is helpful to look at a segment of interaction with a user and describe the messages sent between each of Rea's internal modules.

4.3.1. Interaction 1

The following paragraph records an actual interaction between a user and Rea:

Tim approaches Rea
Rea notices and looks towards him and smiles
Tim says "hello"
Rea responds: "Hello, how can I help you", with a hand wave
Tim says "I'm looking to buy a place near MIT"
Rea glances up and away to keep the turn while "thinking"
Rea says: "I have a house…", with a beat gesture to emphasis the new information "house"
Tim interrupts by beginning to gesture
Rea finishes the current utterance by saying "in Cambridge" and then she gives up the turn.
Tim refines his house request
Rea finishes the house description and then continues.

We will now focus on how the different modules of Rea's architecture contribute to carrying out this interaction. All messages are packaged into a KQML tell-performative as shown in Fig. 3, where the sender and recipient fields contain the names of the modules communicating. For messages that have to do with describing the interaction between the user and Rea, including all messages in this example, the content field contains a frame of type *commact*. The sender and recipient fields of the commact denote where the communicative action originated and who is the intended recipient of it, the value being either *REA* or *USER*, depending on whether the commact is being interpreted or generated by Rea's Decision Module (DM).

The general processing sequence is as follows. IM has some new information about the user's actions and creates a new commact with sender USER and recipient REA. In the input field it places a description of the behaviors detected. The Understanding Module (UM) receives the commact, interprets the behaviors and fills in the prop and intr fields accordingly, sending the commact on to the DM. In reaction to the incoming commact, the DM may construct a new commact, this time with sender REA and recipient USER. After filling in the prop and intr fields, the DM passes the frame on to the GM whose job is to translate the propositional and interactional descriptions into a series of low level behaviors to be placed in the output field. Lastly the AS receives the new commact and using the output field, it coordinates verbal and non-verbal realization.

We will now walk through the specific example given above. As the user comes within a few feet of Rea, a stereoscopic vision system starts to track the user's head and hand movements [30]. Upon receiving this information from the IM, the UM sends the DM an interactional message saying that the user is now present. This makes the system transition into the *UserPresent* state, shown in Fig. 4, sending off to the GM an interactional request for generating an

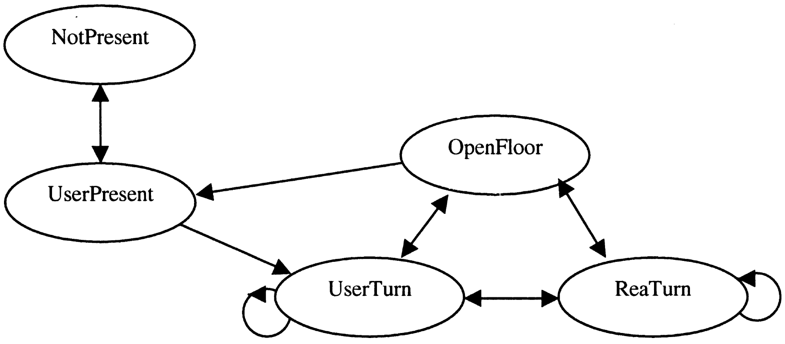

Fig. 4. Rea's conversational states.

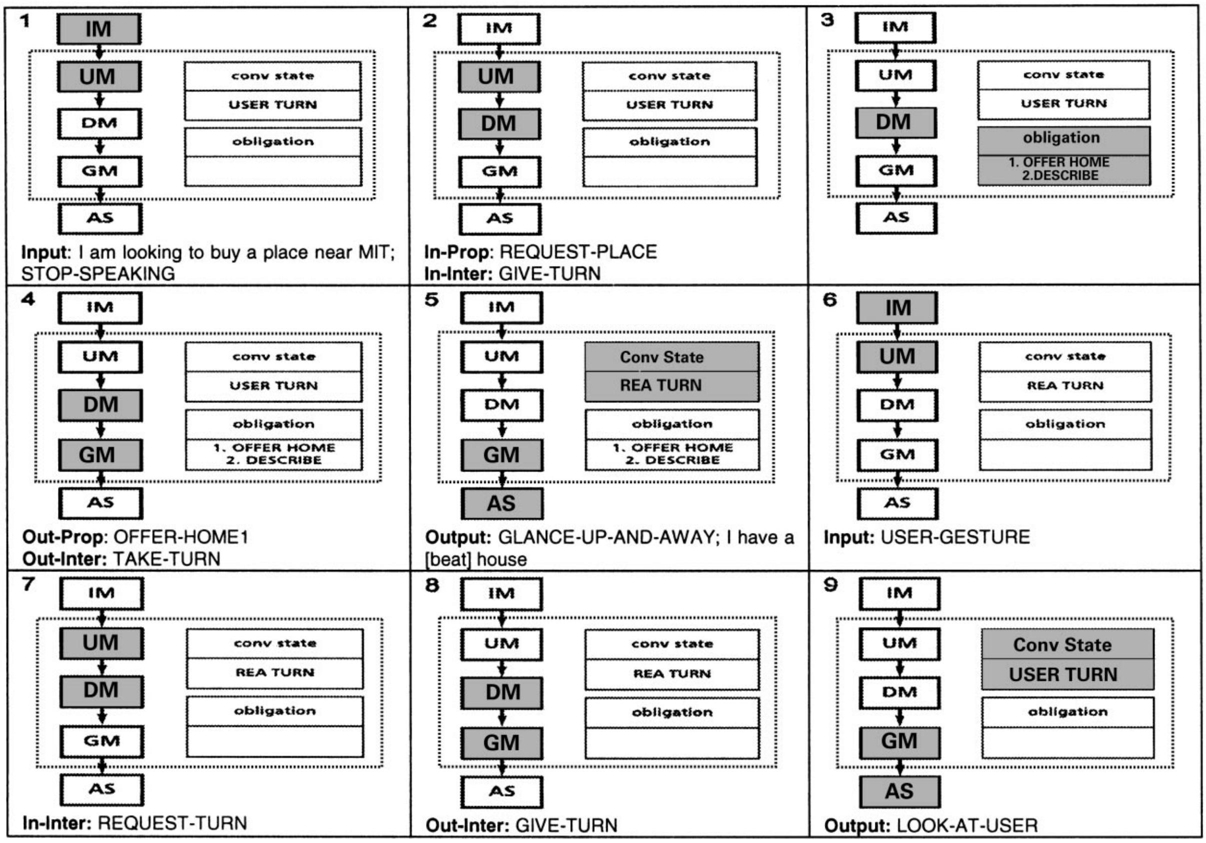

Fig. 5. Sequence showing messages sent between modules, and changes of internal state, during sample interaction 1.

invitation to start a conversation. The GM maps the request to a sequence of behaviors that includes a look towards the user and a smile, to be sent to the AS for execution. When the user responds to the invitation by saying “Hello,” the IM reports the onset of voice to the UM that sends the DM an interactional message saying that the user has now taken the turn. The system transitions into the *UserTurn* state and stays there until the IM delivers the parsed speech content to the UM and the DM has received from the UM an interactional message saying the user has given up the turn, and a propositional message in the form of a speech act, in this case of the type *SA-RITUAL-GREET*. Inside DM this speech act generates an obligation to respond to the greeting. Since a similar *SA-RITUAL-GREET* speech act in return would fulfill that obligation, the DM sends such an act to the GM for execution. The GM breaks the speech act into a hand wave behavior and the spoken utterance “hello, how can I help you” to be realized by the AS. The system is momentarily in a *ReaTurn* state while the speech act is performed, but returns back to an *OpenFloor* state when done.

When the user starts speaking again, the UM produces an interactional message indicating that the user has taken the turn, shifting the system’s state to *UserTurn*. As the user finishes asking “I’m looking to buy a place near MIT” the UM gives the DM the interactional message that the user has given up the turn along with the propositional message that the user performed a *SA-REQUEST-PLACE*. The UM also adds “NearMIT” as an attribute that the place has to have in order to be considered. The DM determines that *HOUSE*1 meets the user’s preferences so the *SA-REQUEST-PLACE* speech act generates an obligation to *OFFER-HOUSE*1. But since this is the first time HOUSE1 is presented to the user, another obligation *DESCRIBE-HOUSE*1 is also generated. Looking at the obligations one at a time, the DM sends off to the GM an *SA-OFFER-HOUSE*1 to fulfill the first one. Along with this propositional message, an interactional message stating that Rea also needs to take the turn is sent to the GM. The GM consults the text and gesture generator for generating an appropriate verbal and gestural expression of the proposition while instructing the AS to glance up and away in an effort to take and keep the turn. The user notices that Rea is planning to speak and does not grab the floor, allowing Rea to stay in a *ReaTurn* state. However, as Rea is delivering the utterance generated by the text and gesture generator, “I have a house”, the user realizes that “near MIT” was perhaps too weak of a constraint and wants to add more detail and therefore spontaneously raises the hands in anticipation of further elaborating on the query. The vision system notices the sudden hand movement and the UM sends a message to the DM saying that the user would like the turn. Gesture is treated as low-priority interrupt, and Rea should finish her current utterance before giving the user the turn, so the DM removes the future obligation to *DESCRIBE-HOUSE*1 but allows the GM and AS to continue executing the current utterance. (Had the user interrupted with speech overlapping Rea's, the DM would also have halted the GM and AS execution, causing Rea to give the user the turn immediately.) When Rea finishes her utterance ("…in Cambridge") she looks at the user in a *UserTurn* state and the user continues. Fig. 5 shows the process just described.

4.3.2. Interaction 2

In this interaction Rea is currently showing Tim a property in her database:

Tim says "show me the kitchen."
Rea shifts the viewpoint of to show the interior of the kitchen, and says "It is a modern kitchen."
Tim says "I like the blue tiles" with a beat gesture on the word 'blue.'
Rea responds by saying "Blue is my favorite color."
Tim says "I like the blue tiles" with a beat gesture on the word 'tiles.'
Rea responds by saying "I love tiles."

In this interaction, the system is already in the *UserTurn* state, and has selected a house to show the user. All of REA's sensors pass messages to the IM, including parsed speech from the speech recognizer, and gesture classifications from the GESTIRP vision system. In this interaction, the IM detects that a user speech event has occurred and that a gesture event has also occurred at approximately the same time. The IM looks at the timestamps of both the speech and gesture and attempts to associate the stroke of the gesture with a particular word in the user's utterance. If the gesture synchronizes with a word, the word is tagged, and the speech and gesture information is bundled into a frame and passed on to the UM for interpretation. In this instance the UM determines that the user's speech act is of type *SA-DECL-USERTASTE* and that a co-occurring beat is present. It then creates a *SA-DECL-USERTASTE* speech act frame and, if the beat stroke co-occurred with a noun or adjective, adds a *USER-EMPHASIS* tag to the frame indicating which word was emphasized. The speech act frame is then passed onto the DM for further processing.

In the DM, the *SA-DECL-USERTASTE* speech act generates an obligation to respond by praising or agreeing with the user's tastes (an obligation common to real estate agents!). If the user emphasizes a color, the DM produces a speech act that involves commenting on the color ("Blue is my favorite color."), whereas if the user emphasizes an object (e.g. tiles) the DM produces a speech act which involves commenting on the object ("I love tiles."). If no user emphasis is detected, a speech act is generated which simply agrees with the user's statement ("Me too!"). The details of turn-taking and message production (via the GM and AS) are the same as for the first sample interaction.

5. Future work

Rea is still a little clumsy in conversation; perhaps not yet your real estate agent of choice. One line of research that we are pursuing is to increase the symmetry of input and output by beginning to sense more conversational protocols in the user, as well as generate more of those protocols in output. To this end, we have begun to develop a sensor to measure head movements and eye gaze using a separate vision system that will estimate what direction the user's face is pointing. This information can be interpreted to infer user turn-taking behaviors, and user backchannel feedback in the form of a nod or headshake. Another line of research attempts to make Rea more truly *responsive* in her conversational behaviors. As we described above, human conversational protocols are co-constructed. For this to be possible in a computational system, Rea must be able to *entrain*, or increasingly adapt her behaviors to be in synchrony with those of the user. To this end, we have begun to implement the ability to engage in *interaction rituals*, such as small talk, so that Rea and the user can gracefully segue into and out of segments of task talk, according to the user's level of comfort [33].

User-testing of the earlier Gandalf system, capable of some of the conversational functions also described here, showed that users relied on the interactional competency of the system to negotiate turn-taking, and that they preferred such a system to another embodied character capable of only facial displays of emotion. In fact, users became so comfortable with Gandalf that they began to overlap their speech with his, which overtaxed his limited speech recognition capabilities [17]. We are currently evaluating Rea to see whether the implementation of a larger set of conversational functions, including error correction and gesture synthesis, allows users to engage in more efficient and fluent interaction with the system. In particular, we are comparing the use of conversational protocols in Rea to similar voice-only conversational protocols in a phone-based dialogue system, so as to draw conclusions about the affordances of embodiment in human–computer conversation.

6. Conclusions

In this paper we have argued that embodied interface agents can provide a qualitative advantage over non-embodied interfaces, if the bodies are used in ways that leverage knowledge of human communicative behavior. We demonstrated our approach with the Rea system. Increasingly capable of making an intelligent content-oriented — or propositional — contribution to the conversation in several modalities, Rea is also sensitive to the regulatory — or interactional — function of verbal and non-verbal conversational behaviors, and is capable of producing regulatory behaviors to improve the interaction.

Acknowledgements

Thanks to the other members of the REA team — in particular David Mellis and Nina Yu — and to Jennifer Smith and Matthew Stone for their contribution to the work and comments on this paper. Thanks to Candy Sidner and several anonymous reviewers for helpful comments that improved the article. Research leading to the preparation of this article was supported by the National Science Foundation (award IIS-9618939), AT&T, Deutsche Telekom, and other generous sponsors of the MIT Media Lab.

References

[1] E. Boyle, A. Anderson, A. Newlands, The effects of visibility in a cooperative problem solving task, Language and Speech 37 (1) (1994) 1–20.

[2] J. Beskow, S. McGlashan, Olga: a conversational agent with gestures. IJCAI'97, 1997.

[3] A. Takeuchi, K. Nagao, Communicative facial displays as a new conversational modality. InterCHI'93, ACM Press, Amsterdam, 1993.

[4] R. Brooks, A Robust Layered Control System for a Mobile Robot, MIT AI Lab, Cambridge, MA, 1985.

[5] J.C. Lester et al., Cosmo: a life-like animated pedagogical agent with deictic believability. IJCAI'97, 1997.

[6] J. Rickel, W.L. Johnson, Animated agents for procedural training in virtual reality: perception, cognition, and motor control, Applied Artificial Intelligence 13 (4/5) (1999) 343–382.

[7] J. Cassell et al., Animated conversation: rule-based generation of facial expression, gesture and spoken intonation for multiple conversational agents. Siggraph '94, ACM Press, Orlando, 1994.

[8] K. Thorisson, Communicative Humanoids: A Computational Model of Psychosocial Dialogue Skills. MIT Media Laboratory Ph.D. thesis, MIT, Cambridge, MA, 1996.

[9] S. Prevost et al., Face-to-face interfaces. CHI'99, ACM Press, 1999.

[10] W. Wahlster et al. Designing illustrated texts. EACL'91, 1991.

[11] N. Green et al. A media-independent content language for integrated text and graphics generation. Workshop on Content Visualization and Intermedia Representations at COLING and ACL'98, 1998.

[12] S. Feiner, K. McKeown, Automating the generation of coordinated multimedia explanations, IEEE Computer 24 (10) (1991) 33–41.

[13] T. Koda, P. Maes, Agents with faces: the effects of personification of agents. Fifth IEEE International Workshop on Robot and Human Communication, 1996.

[14] A. Takeuchi, T. Naito, Situated facial displays: towards social interaction. Human Factors in Computing Systems: CHI'95, ACM Press, 1995.

[15] S. Kiesler, L. Sproull, Social human–computer interaction, in: B. Friedman (Ed.), Human Values and the Design of Computer Technology, 199, CSLI Publications, Stanford, CA, 1997, p. 191.

[16] E. Andre, T. Rist, J. Muller, Integrating reactive and scripted behaviors in a life-like presentation agent. Proceedings of AGENTS'98, pp. 261–268, 1998.

[17] J. Cassell, K.R. Thorisson, The power of a nod and a glance: envelope vs. emotional feedback in animated conversational agents, Applied Artificial Intelligence 13 (1999) 519–538.

[18] N. Chovil, Discourse-oriented facial displays in conversation, Research on Language and Social Interaction 25 (1992) 163–194.

[19] A. Kendon, Conducting Interaction: Patterns of behavior in Focused Encounters, Cambridge University Press, New York, 1990.

[20] S. Duncan, On the structure of speaker–auditor interaction during speaking turns, Language in Society 3 (1974) 161–180.

[21] J. Cappella, C. Pelachaud, Rules for responsive robots: using human interaction to build virtual interaction, in: Reis, Fitzpatrick, Vangelisti (Eds.), Stability and Change in Relationships, in preparation.

[22] D. McNeill, Hand and Mind: What Gestures Reveal about Thought, The University of Chicago Press, Chicago, IL, 1992.

[23] J. Cassell, et al., Human conversation as a system framework: designing embodied conversational agents, in: J. Cassell, et al. (Eds.), Embodied Conversational Agents, MIT Press, Cambridge, MA, 2000.

[24] J. Cassell et al., Embodiment in conversational interfaces: Rea. CHI '99, ACM Press, Pittsburgh, PA, 1999.

[25] M. Stone, Modality in Dialogue: Planning, Pragmatics, and Computation, University of Pennsylvania, 1998.

[26] J. Cassell, M. Stone, H. Yan, Coordination and context-dependence in the generation of embodied conversation. INLG 2000, Mitzpe Ramon, Israel, 2000.

[27] T. Finin et al., Specification of the KQML Agent-Communication Language, 1994.

[28] CLIPS Reference Manual Version 6.0, Software Technology Branch, Lyndon B. Johnson Space Center, Houston, TX.

[29] H. Yan, Paired Speech and Gesture Generation in Embodied Conversational Agents, MIT Media Laboratory MA thesis, MIT, Cambridge, MA, 2000.

[30] A. Azarbayejani, C. Wren, A. Pentland, Real-time 3-D Tracking of the Human Body. IMAGE'COM, Bordeaux, France, 1996.

[31] L. Campbell, Visual Classification of Discourse Gestures for Gesture Understanding, MIT Media Laboratory PhD thesis, MIT, Cambridge, MA, 2000.

[32] L.R. Rabiner, A tutorial on hidden Markov models and selected applications in speech recognition, Proceedings of the IEEE 77 (2) (1989) 257–286.

[33] J. Cassell, T. Bickmore, External manifestations of trustworthiness, Communications of the ACM (2000) 12/00.