Embodiment in Conversational Interfaces: Rea*

J. Cassell, T. Bickmore, M. Billinghurst, L. Campbell, K. Chang, H. Vilhjálmsson, H. Yan
Gesture and Narrative Language Group
MIT Media Laboratory
E15-315
20 Ames St, Cambridge, Massachusetts
+1 617 253 4899
{justine, bickmore, markb, elwin, tetrion, hannes, yanhao}@media.mit.edu

ABSTRACT
In this paper, we argue for *embodied conversational characters* as the logical extension of the metaphor of human – computer interaction as a conversation. We argue that the only way to fully model the richness of human face-to-face communication is to rely on conversational analysis that describes sets of conversational behaviors as fulfilling conversational functions, both interactional and propositional. We demonstrate how to implement this approach in Rea, an embodied conversational agent that is capable of both multimodal input understanding and output generation in a limited application domain. Rea supports both social and task-oriented dialogue. We discuss issues that need to be addressed in creating embodied conversational agents, and describe the architecture of the Rea interface.

Keywords
Conversational Characters, Multimodal Input, Intelligent Agents, Multimodal Output

INTRODUCTION
The metaphor of face-to-face conversation has been successfully applied to human-interface design for quite some time. One of the early descriptions of this metaphor gave a list of features of face-to-face conversation that could be fruitfully applied to HCI, including mixed initiative, non-verbal communication, sense of presence, rules for transfer of control, and so forth [1]. However, although these features have gained widespread recognition, human – computer conversation has never become more than a metaphor. That is, designers have not taken the metaphor seriously in such a way as to design a computer that could hold up its end of the conversation.

In the current paper we argue that while this metaphor has been useful to HCI, its use to date has been just that; a metaphor. We believe that interfaces that are truly conversational have the promise of being more intuitive to learn, more resistant to communication breakdown, and more functional in high *noise* environments. Therefore, we propose to leverage the full breadth and power of human conversational competency by imbuing the computer with all of the conversational skills that humans have; to whit, the ability to use the face, hands, and melody of the voice to regulate the process of conversation, as well as the ability to use verbal and nonverbal means to contribute content to the ongoing conversation.

In addition, we argue that the only way to accomplish such a goal of *embodying* the interface is to implement a model of *conversational function*. This means that particular conversational behaviors (such as head nods and expressions of agreement) are generated and understood in terms of the functions that they fulfill in the ongoing conversation (such as 'take turn', 'contribute new information').

To provide a practical example of this approach, we present Rea, an embodied conversational agent whose verbal and nonverbal behaviors are designed in terms of conversational functions. Rea is not designed with the metaphor of the interface as a conversation, but actually implements the social, linguistic, and psychological conventions of conversation. Rea differs from other dialogue systems, and other conversational agents in three ways:

* Support for this research provided by NSF-STIMULATE IRI-9618939.

- Rea has a human-like body, and uses her body in human-like ways during the conversation. That is, she uses eye gaze, body posture, hand gestures, and facial displays to organize and regulate the conversation.
- The underlying approach to conversational understanding and generation in Rea is based on discourse functions. Thus, each of the users' inputs are interpreted in terms of their conversational function and responses are generated according to the desired function to be fulfilled. Such models have been described for other conversational systems: for example Brennan and Hulteen describe a general framework for applying conversational theory to speech interfaces [7]. Our work extends this by developing a conversational model that that relies on the function of non-verbal behaviors as well as speech, and that makes explicit the interactional and propositional contribution of these conversational behaviors.
- Rea is being designed to respond to visual, audio and speech cues normally used in face to face conversation, such as speech, shifts in gaze, gesture, and non-speech audio (feedback sounds). She is being designed to generate these cues, ensuring a full symmetry between input and output modalities. This is a step towards enabling Rea to participate on more of an equal footing with the user in a human-computer conversation.

Developing an embodied conversational agent is a complex endeavor that draws on many fields. We begin this paper by describing several motivations for building embodied conversational agents. We then review past work in relevant HCI areas, and in several theories of conversation. Examination of these theories leads us to believe that a conversational function approach may be the most appropriate for a conversational agent. We then present Rea, and describe how we have begun to implement conversational function in an embodied interface agent.

Motivation

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

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

There are a number of motivations for developing interfaces with these attributes, including:

Intuitiveness. Conversation is an intrinsically human skill that is learned over years of development and is practiced daily. Conversational interfaces provide an intuitive paradigm for interaction, since the user is not required to learn new skills.

Redundancy and Modality Switching: Embodied conversational interfaces support redundancy and complementarity between input modes. This allows the user and system to increase reliability by conveying information in more than one modality, and to increase expressiveness by using each modality for the type of expression it is most suited to.

The Social Nature of the Interaction. Whether or not computers look human, people attribute to them human-like properties such as friendliness, or cooperativeness [22]. An embodied conversational interface can take advantage of this and prompt the user to naturally engage the computer in human-like conversation. If the interface is well-designed to reply to such conversation, the interaction may be improved

As we shall show in the next section, there has been significant research in the areas of conversational analysis and multimodal interfaces. However there has been little work in the recognition and use of conversational cues for conversational interfaces, or the development of computational conversational models that support non-speech input and output. A prime motivation for our work is the belief that effective embodied conversational interfaces cannot be built without an understanding of verbal and non-verbal conversational cues, and their function in conversation.

RELATED WORK

There are many challenges that must be overcome before embodied conversational interfaces reach their full potential. These range from low-level issues such as capturing user input to high level problems such as agent planning and dialogue generation. In this section we review related work in three areas; multimodal interfaces, models of conversation, and conversational agent interfaces.

Multimodal Interfaces

Embodied conversational agents are similar to *multimodal* systems in that information from several modalities must be integrated into one representation of speaker intention. One of the first multimodal systems was *Put-That-There*, developed by Bolt, Schmandt and their colleagues [5]. *Put That There* used speech recognition and a six-degree-of-freedom space sensing device to gather input from a user's speech and the location of a cursor on a wall-sized display, allowing for simple deictic reference to visible entities. More recently, several systems have built on this early work. Koons allowed users to maneuver around a two-dimensional map using spoken commands, deictic hand gestures, and eye gaze [16]. In this system, nested frames were employed to gather and combine information from the

different modalities. As in Put-that-There, speech drove the analysis of the gesture: if information is *missing* from speech, then the system will search for the missing information in the gestures and/or gaze. Time stamps unite the actions in the different modalities into a coherent picture. Wahlster used a similar method, also depending on the linguistic input to guide the interpretation of the other modalities [27]. Bolt and Herranz described a system that allows a user to manipulate graphics with two-handed semi-iconic gesture [6]. Using a cutoff point and time stamping, motions can be selected that relate to the intended movement mentioned in speech. Sparrell used a scheme based on stop-motion analysis: whenever there is a significant stop or slowdown in the motion of the user's hand, then the preceding motion segment is grouped and analyzed for features such as finger posture and hand position [23]. In all of these systems interpretation is not carried out until the user has finished the utterance.

Johnston describes an approach to understanding of user input based on unification across grammars that can express input from multiple modalities[14]. While the system does treat modalities equally (vs. filling in utterance-based forms) it is still based on a mapping between combinations of specific gestures and utterances on the one hand, and user intentions (commands) on the other hand. In addition, all behaviors are treated as propositional -- none of them control the envelope of the user-computer interaction.

Although these works are primarily command-based rather than conversational, there are some lessons we can learn from them, such as the importance of modeling the user and developing interfaces which use existing deeply ingrained conversational behaviors [21]. They also highlight areas of potential difficulty, such as the fact that humans do not naturally use gesture according to a grammar with standards of form or function, and the problem of recognition errors in speech and gesture.

Missing from these systems is a concept of non-verbal function with respect to conversational function. That is, in the systems reviewed thus far, there is no discourse structure over the sentence (no notion of "speaking turn" or "information structure" [24]). Therefore the role of gesture and facial expression cannot be analyzed at more than a sentence-constituent-replacement level. Gestures are only analyzed as support for referring expressions (gestures provide the referent for demonstratives such as “that”). What is needed is a discourse structure that can take into account why one uses a verbal or nonverbal device in a particular situation, and a conversational structure that can account for how non-verbal behaviors function in conversation regulation – such as turn-taking – as well as conversational content.

Conversational Models

Even though conversation is considered an orderly event, governed by rules, no two conversations look exactly the same and the set of behaviors exhibited differs from person to person and from conversation to conversation. Therefore to successfully build a model of how conversation works, one can not refer to surface features, or *conversational behaviors* alone. Instead, the emphasis has to be on identifying the fundamental phases and high level structural elements that make up a conversation. These elements are then described in terms of their role or *function* in the exchange. Typical discourse functions include *conversation invitation*, *turn taking*, *providing feedback*, *contrast and emphasis*, and *breaking away* [10][15].

It is important to realize that each of these functions can be realized in several different manners. The form we give to a particular discourse function depends on, among other things, current availability of modalities, type of conversation, cultural patterns and personal style. For example to emphasize a point one can strike a fist into the table, nod the head, raise the eyebrows, apply rising intonation or construct some combination of these. In a different context these behaviors may carry a different meaning, for example a head nod can indicate back-channel feedback or a salutation rather than emphasis.

Despite the fact that different behaviors may fulfill the same function, it is striking the extent to which such non-verbal behaviors coordinate and regulate conversation. It is clear that through gaze, eyebrow raises and head nods both speakers and listeners collaborate in the construction of synchronized turns, and efficient conversation. In this way, these non-verbal behaviors participate in *grounding* the conversation [11], and fill the functions that Brennan & Hulteen (1995) suggest are needed for more robust speech interfaces [7].

An important aspect of the grounding of a conversation is evidence of understanding [11]. This includes means such as paraverbals (“huh?”, “Uh-huh!”) and other back channel feedback. A conversational model that uses both positive and negative feedback enables an agent to recognize a misunderstanding and initiate the appropriate repair mechanisms.

To further clarify these types of roles fulfilled by discourse behaviors, the contribution to the conversation can be divided into *propositional information* and *interactional information*. Propositional information corresponds to the content of the conversation. This includes meaningful speech as well as hand gestures and intonation used to complement or elaborate upon the speech content (gestures that indicate size in the sentence “it was *this* big” or rising intonation that indicates a question with the sentence “you went to the store”). Interactional information consists of cues that regulate the conversational process and includes a range of nonverbal behaviors (quick head nods to indicate that one is following) 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, while propositional functions shape the actual content.

Although the way in which conversation incorporates speech and other movements of the body has been studied for some time, there have been few attempts by the engineering community to develop embodied computer interfaces based on this understanding. On the contrary, embodied conversational characters have, for the most part been built with hardwired associations between verbal and non-verbal conversational behaviors, without a clear flexible notion of conversational function underlying those behaviors. In interfaces of this sort, there is no possibility for one modality to take over for another, or the two modalities to autonomously generate complementary information. Thus, a primary goal of our work is to map multiple modalities onto discourse functions, both for input and output. Input events in different modalities may be mapped onto the same discourse function, while in different conversational states the same function may lead to different conversational behaviors, based on state, as well as the availability of input and output modalities.

Embodied Conversational Interfaces

Other researchers have built embodied conversational agents, with varying degrees of conversational ability. Ball et al. are building an embodied conversational interface that will eventually integrate spoken language input, a conversational dialogue manager, reactive 3D animation, and recorded speech output [3]. Each successive iteration of their computer character has made significant strides in the use of these different aspects of an embodied dialogue system. Although their current system uses a tightly constrained grammar for NLP and a small set of pre-recorded utterances that their character can utter, it is expected that their system will become more generative in the near future. Their embodiment takes the form of a parrot. This has allowed them to simulate gross "wing gestures" (such as cupping a wing to one ear when the parrot has not understood a user's request) and facial displays (scrunched brows as the parrot finds an answer to a question). The parrot's output, however, is represented as a set of conversational behaviors, rather than a set of conversational functions. Therefore, modalities cannot share the expressive load, or pick up the slack for one another in case of noise, or in the case of one modality not being available. Nor can any of the modalities regulate a conversation with the user, since user interactional behaviors cannot be perceived or responded to.

Loyall and Bates build engaging characters that allow the viewer to suspend disbelief long enough to interact in interesting ways with the character, or to be engaged by the character's interactions with another computer character [17]. Associating natural language with non-verbal behaviors is one way of giving their characters believability. In our work, the causality is somewhat the opposite: we build characters that are believable enough to allow the use of language to be human-like. That is, we believe that the use of gesture and facial displays does make the characters life-like and therefore believable, but these communicative behaviors also play integral roles in enriching the dialogue, and regulating the process of the conversation. It is these latter functions that are most important to us. In addition, like Ball et al., the Oz group has chosen a very non-human computer character—Woggles, which look like marbles with eyes. Researchers such as Ball and Bates argue that humanoid characters raise users' expectations beyond what can be sustained by interactive systems and therefore should be avoided. We argue the opposite, that humanoid interface agents do indeed raise users' expectations . . . up to what they expect from humans, and therefore lower their difficulty in interacting with the computer, which is otherwise for them an unfamiliar interlocutor (as is a marble, as well).

Noma & Badler have created a virtual human weatherman, based on the *Jack* human figure animation system [20]. In order to allow the weatherman to gesture, they assembled a library of presentation gestures culled from books on public speaking, and allowed authors to embed those gestures as commands in text that will be sent to a speech-to text system. This is a useful step toward the creation of presentation agents of all sorts, but does not deal with the autonomous generation of non-verbal behaviors in conjunction with speech. Other efforts along these lines include André et al. [1] and Beskow and McGlashan [4].

The work of Thórisson provides a good first example of how discourse and non-verbal function might be paired in a conversational multimodal interface [26]. In this work the main emphasis was the development of a multi-layer multimodal architecture that could support fluid face-to-face dialogue between a human and graphical agent. The agent, Gandalf, was capable of discussing a graphical model of the solar system in an educational application. Gandalf recognized and displayed interactional information such as head orientation, simple pointing and beat gestures and canned speech events. In this way it was able to perceive and generate turn-taking and back channel behaviors that lead to a more natural conversational interaction.

However, Gandalf had limited ability to recognize and generate propositional information, such as providing correct intonation for speech emphasis on speech output, or a content-carrying gesture with speech. "Animated Conversation" [8] was a system that automatically generated context-appropriate gestures, facial movements and intonational patterns. In this case the challenge was to generate 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. Since there was no interaction with a real user, the interactional information was very limited, and not reactive

(although some interactional types of face and head movements, such as nods, were generated).

Rea is an attempt to develop an agent with both propositional and interactional understanding and generation, which can interact with the user in real time. As such it combines elements of the Gandalf and Animated Agents projects into a single interface and moves towards overcoming the limitations of each. In the next section we describe interaction with the Rea agent and its implementation.

REA: AN EMBODIED CONVERSATIONAL AGENT

The Rea Interface

Rea ("Real Estate Agent") is a computer generated humanoid that has an articulated graphical body, can sense the user passively through cameras and audio input, and is capable of speech with intonation, facial display, and gestural output (Figure 1). The system currently consists of a large 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 graphics and conversation engine of Rea, while several other computers manage the speech recognition and generation and image processing.

Figure 1. User Interacting with Rea

A Sample Interaction

Rea's domain of expertise is real estate and she acts as a real estate agent showing users the features of various models of houses that appear on-screen behind her. The following is a excerpt from a sample interaction:

Lee approaches the projection screen. Rea is currently turned side on and is idly gazing about. As the user moves within range of the cameras, Rea turns to face him and says "Hello, my name is Rea, what's your name?"

"Lee"

"Hello Lee would you like to see a house?" Rea says with rising intonation at the end of the question.

"That would be great"

A picture of a house appears on-screen behind Rea.

"This is a nice Victorian on a large lot" Rea says gesturing towards the house. "It has two bedrooms and a large kitchen with.."

"Wait, tell me about the bedrooms" Lee says interrupting Rea by looking at her and gesturing with his hands while speaking.

"The master bedroom is furnished with a four poster bed, while the smaller room could be used for a children's bedroom or guest room. Do you want to see the master bedroom?".

"Sure, show me the master bedroom". Lee says, overlapping with Rea.

"I'm sorry, I didn't quite catch that, can you please repeat what you said", Rea say.

And the house tour continues…

Rea is designed to conduct a mixed initiative conversation, pursuing the goal of describing the features of a house that fits the user's requirements while also responding to the users' verbal and non-verbal input that may lead in new directions. 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 repair when she misunderstands what the user says, and can generate combined voice and gestural output. For the moment, Rea's responses are generated from an Eliza-like engine that mirrors features of the user's last utterance [28], but efforts are currently underway to implement an incremental natural language and gesture generation engine, along the lines of [8].

In order to carry on natural conversation of this sort, Rea uses a conversational model that supports multimodal input and output as constituents of conversational functions. That is, input and output is interpreted and generated based on the discourse functions it serves. The multimodal conversational model and the underlying Rea architecture are discussed in the next sections.

Implementation

While Rea is capable of understanding speech, and making reasonable contributions to an ongoing conversation about realty, to date our primary effort has been in the interactional component of the conversational model. This component manages several discourse functions. The functions currently being managed are:

- Acknowledgment of user's presence - by posture, turning to face the user;
- Feedback function - Rea gives feedback in several modalities: she may nod her head or emit a paraverbal (e.g. "mmhmm") or a short statement such as "I see" in response to short pauses in the user's speech; she raises her eyebrows to indicate partial understanding of a phrase or sentence.

- Turntaking function – Rea tracks who has the speaking turn, and only speaks when she 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 she will interpret this as expression of a desire to speak [15], and therefore halt her remarks at the nearest sentence boundary. Finally, at the end of her speaking turn she turns to face the user to indicate the end of her turn.

Other functions have both interactional and propositional content. For example:

- Greeting and Farewell functions - Rea speaks and gestures when greeting and saying goodbye.
- Emphasis function - people may emphasize particular linguistic items by prosodic means (pitch accents) or by accompanying the word with a beat gesture (short formless wave of the hand). Recognizing emphasis is important for determining which part of the utterance is key to the discourse. For example, the user may say "I'd like granite floor tiles," to which Rea can reply "granite is a good choice here;" or the user might say "I'd like granite floor tiles," where Rea can reply "tile would go well here." We are developing a gesture classification system to detect the 'beat' gestures that often indicate emphasis. On the output side, we plan to allow Rea to generate emphasis using either modality.

These conversational functions are realized as conversational behaviors. For turn taking, for example, the specifics are as follows:

If Rea has the turn and is speaking and the user begins to gesture, this is interpreted as the user *wanting turn* function. If Rea has the turn and is speaking and the user begins to speak, this is interpreted as the user *taking turn* function. If the user is speaking and s/he pauses for less than 500 msec., this is interpreted as the *wanting feedback* function. If the user is speaking and issues a declarative sentence and stops speaking and gesturing, or says an imperative or interrogative phrase, their input is interpreted as a *giving turn* function. Finally, if the user has the turn and continues gesturing after having finished uttering a declarative sentence, or if s/he begins another phrase after having uttered a declarative sentence, with a pause of less than 500 msec, this is interpreted as a *holding turn* function. This approach is summarized in Table 1.

Thus, speech may convey different interactional information; it may be interpreted as taking turn, giving turn, or holding turn depending on the conversational state and what is conveyed by the other modalities.

A similar approach is taken for generation of conversational behaviors. Rea generates speech, gesture and facial expressions based on the current conversational state and the conversational function she is trying to convey. For example, when the user first approaches Rea ("User Present" state), she signals her openness to engage in conversation by looking at the user, smiling, and/or tossing her head. When conversational turn-taking begins, she orients her body to face the user at a 45 degree angle. When the user is speaking and Rea wants the turn she looks at the user and utters a paraverbal ("umm"). When Rea is finished speaking and ready to give the turn back to the user she looks at the user, drops her hands out of gesture space and raises her eyebrows in expectation. Table 2 summarizes Rea's current interactional output behaviors.

State	User Input	Input Function
Rea speaking	Gesture	Wanting turn
	Speech	Taking turn
User speaking	Pause of <500 msec.	Wanting feedback
	Imperative phrase	Giving turn
	Interrogative phrase	Giving turn
	Declarative phrase & pause >500 msec. & no gesture	Giving turn
	Declarative phrase & long gesture or pause	Holding turn

Table 1. Functional interpretation of turn taking input

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

Table 2. Output Functions

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, she can choose any of several modalities based on what is appropriate at the moment.

Architecture

Figure 2 shows the modules of the Rea architecture. The three points that differentiate Rea from other embodied conversational agents are mirrored in the organization of the system architecture.

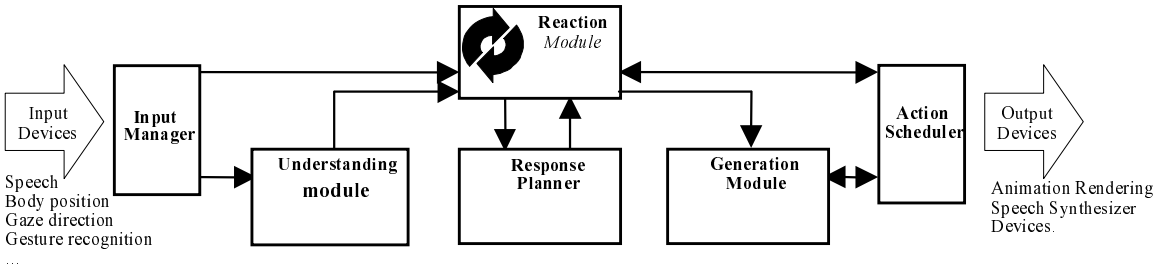

Figure 2: The Rea Software Architecture

- 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 [13].
- The KQML 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.
- 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 (the understanding, response planner, and generation modules), and moves to the periphery decisions made in terms of behaviors (the input manager and action scheduler).

In addition, a distinction is drawn between *reactive* and *deliberative* communicative actions [26]. The Input Manager and Action Scheduler interact with external devices and together with the Reaction Module respond immediately (under 500 msec.) to user input or system commands. Performing head nods when the user pauses briefly is an example of a reactive conversational behavior. The other modules are more "deliberative" in nature and perform non-trivial inferencing actions that can take multiple real-time cycles to complete. These modules are written in C++ and CLIPS, a rule-based expert system language [12].

Input Manager

The input manager currently supports three types of input:

- *Gesture Input:* STIVE vision software produces 3D position and orientation of the head and hands[2].
- *Audio Input:* A simple audio processing routine detects the onset and cessation of speech.
- *Grammar Based Speech Recognition:* IBM ViaVoice returns text from a set of phrases defined by a grammar.

In all cases the features sent to the Input Manager are time stamped with start and end times in milliseconds. The various computers are synchronized to within a few milliseconds of each other. This synchronization is key for associating verbal and nonverbal behaviors. Latency in input devices can have a significant impact on the functioning of the system, since delays of milliseconds can have significant meaning in conversation. For example, if Rea delays before giving a "yes" response it can be interpreted by the user as indecision. Thus, our goal is to minimize input device and processing latencies wherever possible.

Low level gesture and audio detection events are sent to the reaction module straight away. These events are also stored in a buffer so that when recognized speech arrives a high-level multimodal KQML frame can be created containing mixed speech, audio and gesture events. This is sent to the understanding module for interpretation.

Understanding Module

The Understanding Module fuses all input modalities into a coherent understanding of what the user might be doing based on the current conversational state.

Reaction Module

The Reaction Module is responsible for the "action selection" component of the architecture, which determines at each moment in time what the character should be doing.

Response Planner Module

The Response Planner formulates sequences of actions, some or all of which will need to be executed during future execution cycles, to carry out desired communicative or task goals.

Generation Module

The Generation Module realizes a complex action request from the Reasoning Module by producing one or more coordinated primitive actions (such as speech or gesture generation, or facial expression), sending them to the Action Scheduler, and monitoring their execution.

Action Scheduling Module

The Action Scheduling Module is the "Motor controller" for the character, responsible for coordinating action at the lowest level. It takes multiple action requests from multiple requestors (i.e. the Reaction and Generation Modules) and attempts to carry them out.

Conclusion

User-testing of Gandalf, 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 emotional expression. However, Gandalf did not handle repairs gracefully, and users were comparatively more disfluent when using the system [9]. Our next step is to test Rea to see whether the current mixture of interactional and propositional conversational functions, including turn-taking and repair, allow users to engage in more efficient and fluent interaction with the system.

The functional approach provides abstraction that not only serves theoretical goals but also gives important leverage for multi-cultural scalability. The inner workings of the system deal with a set of universal conversational functions while the outer modules, both on the input and output side, are responsible for mapping them onto largely culture-specific surface behaviors. The architecture allows us to treat the mappings as an easily exchangeable part in the form of a specification file.

In this paper we have argued that embodied conversational agents are a logical and needed extension to the conversational metaphor of human – computer interaction. We argue, however, that embodiment needs to be based on an understanding of conversational function, rather than a additive – and ad hoc -- model of the relationship between nonverbal modalities and verbal conversational behaviors.

We demonstrated our approach with the Rea system. Increasingly capable of making an intelligent content-oriented – or *propositional* – contribution to the conversation, 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 by helping the user remain aware of the state of the conversation. Rea is an embodied conversational agent who can hold up her end of the conversation.

REFERENCES

1. Andre, E., Rist, T., Mueller, J. Integrating Reactive and Scripted Behaviors in a Life-Like Presentation Agent. In Proceedings of Autonomous Agents 98, (Minneapolis/St. Paul, May 1998), ACM Press,
2. Azarbayejani, A., Wren, C. and Pentland A. Real-time 3-D tracking of the human body. In *Proceedings of IMAGE'COM 96,* (Bordeaux, France, May 1996).
3. Ball, G., Ling, D., Kurlander, D., Miller, D., Pugh, D., Skelly, T., Stankosky, A., Thiel, D., Van Dantzich, M. and T. Wax. Lifelike computer characters: the persona project at Microsoft Research. In *Software Agents,* J. M. Bradshaw (ed.), MIT Press, Cambridge, MA, 1997.
4. Beskow, J. and McGlashan, S. Olga - A Conversational Agent with Gestures, In *Proceedings of the IJCAI'97 workshop on Animated Interface Agents - Making them Intelligent*, (Nagoya, Japan, August 1997), Morgan-Kaufmann Publishers, San Francisco.
5. Bolt, R.A. Put-that-there: voice and gesture at the graphics interface. *Computer Graphics*, 14(3), 1980, 262-270.
6. Bolt, R.A. and Herranz, E. Two-handed gesture in multi-modal natural dialog. In *Proceedings of UIST `92, Fifth Annual Symposium on User Interface Software and Technology*, (Monterey, CA, November 1992). ACM Press, 7-14.
7. Brennan, S.E. and Hulteen, E.A. Interaction and Feedback in a Spoken Language System. *Knowledge-Based Systems,* 8,2 (April-June 1995), 143-151.
8. Cassell, J., Pelachaud, C., Badler, N.I., Steedman, M., Achorn, B., Beckett, T., Douville, B., Prevost, S. and Stone, M. Animated conversation: rule-based generation of facial display, gesture and spoken intonation for multiple conversational agents. *Computer Graphics (SIGGRAPH '94 Proceedings)*, 28(4): 413-420.
9. Cassell, J. and Thórisson, K. The Power of a Nod and a Glance: Envelope vs. Emotional Feedback in Animated Conversational Agents. *Journal of Applied Artificial Intelligence*, in press.
10. Cassell, J., Torres, O. and Prevost, S. Turn taking vs. Discourse Structure: how best to model multimodal conversation. In Wilks (ed.) *Machine Conversations.* Kluwer, The Hague, 1998.
11. Clark, H.H. and Brennan, S.E. Grounding in Communication. In *Shared Cognition: Thinking as Social Practice,* J. Levine, L.B. Resnick and S.D. Behrend, (eds.). APA Books, Washington, D.C, 1991.
12. CLIPS Reference Manual Version 6.0. *Technical Report,* Number JSC-25012, Software Technology Branch, Lyndon B. Johnson Space Center, Houston, TX, 1994.
13. Finin, T., Fritzson, R. KQML as an Agent Communication Language. In *The Proceedings of the Third International Conference on Information and Knowledge Management* (CIKM'94), ACM Press, November 1994.
14. Johnston, M., Cohen, P. R., McGee, D., Oviatt, S. L., Pittman, J. A. and Smith, I. Unification-based multimodal integration. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics,* (Madrid, Spain, 1997).
15. Kendon, A. The negotiation of context in face-to-face interaction. In A. Duranti and C. Goodwin (eds.), *Rethinking context: language as interactive phenomenon.* Cambridge University Press. NY, 1990.

16. Koons, D.B. Sparrell, C.J. and Thorisson, K.R. Integrating simultaneous input from speech, gaze and hand gestures. In *Intelligent Multi-Media Interfaces* M.T. Maybury (Ed.), AAAI Press/MIT Press, 1993.

17. Loyall, A. and Bates, J. Personality-rich believable agents that use language. In Proceedings of Agents '97 (Marina del Rey, CA, February 1997), ACM Press,

18. Nagao, K. and Takeuchi, A. Social interaction: multimodal conversation with social agents. *Proceedings of the 12th National Conference on Artificial Intelligence (AAAI-94),* (Seattle, WA, August 1994), AAAI Press/MIT Press, vol. 1, 22-28.

19. Nickerson, R.S. On Conversational Interaction with Computers. In *User Oriented Design of Interactive Graphics Systems: Proceedings of the ACM SIGGRAPH Workshop* (1976), ACM Press, 681-683.

20. Noma, T. and Badler, N. (1997). A virtual human presenter. In *Proceedings of the IJCAI'97 workshop on Animated Interface Agents - Making them Intelligent*, (Nagoya, Japan, August, 1997), Morgan-Kaufmann Publishers, San Francisco.

21. Oviatt, S.L. User-Centered Modeling for Spoken Language and Multimodal Interfaces, *IEEE Multimedia,* 3, 4, (Winter 1996), 26-35.

22. Reeves, B. and Nass, C. *The Media Equation: How People Treat Computers, Television, and New Media Like Real People and Places.* Cambridge University Press, 1996.

23. Sparrell, C. J. *Coverbal Iconic Gestures in Human-Computer Interaction.* S.M. Thesis, MIT Media Arts and Sciences Section, 1993.

24. Steedman, M. Structure and intonation. *Language*, 1991, 67(2), 190-296.

25. Takeuchi, A. and Nagao, K. Communicative facial displays as a new conversational modality. In *Proceedings of InterCHI 93,* (Amsterdam, Netherlands April 1993), 187-193.

26. Thórisson, K. R. *Communicative Humanoids: A Computational Model of Psychosocial Dialogue Skills.* PhD Thesis, MIT Media Laboratory, 1996.

27. Wahlster, W., André, E., Graf, W. and Rist, T. Designing illustrated texts. In *Proceedings of the 5th EACL* (Berlin, Germany, April 1991), 8-14

28. Weizenbaum, J. Eliza. A computer program for the study of natural language communication between man and machine. *Communications of the ACM*, 1966, 9, 26-45.