(in press) Spoken Dialogue Systems
S. Luperfoy (ed.)
MIT Press

Embodied Conversation: Integrating Face and Gesture into Automatic Spoken Dialogue Systems[†]

Justine Cassell

MIT Media Laboratory

Introduction

In this chapter I'm going to discuss the issues that arise when we design automatic spoken dialogue systems that can use not only voice, but also facial and head movements and hand gestures to communicate with humans. For the most part I will concentrate on the generation side of the problem that is, building systems that can speak, move their faces and heads and make hand gestures. As with most aspects of spoken dialogue, however, generation is no good without comprehension, and so I will also briefly discuss some of the issues involved in building systems that understand non-verbal communicative behaviors.

[†] Thanks to the Gesture Jack team, especially Catherine Pelachaud, Norm Badler, Mark Steedman, and Matthew Stone, and to the Gesture and Narrative Language team, especially Tim Bickmore, Scott Prevost, Kris Thorisson, Hannes Vilhjálmsson, Sola Grantham and Erin Panttaja, for the major role they have played in clarifying the issues, and creating the systems.

Why would it even occur to us to add these non-verbal modalities to systems that have always been called *spoken* dialogue systems (rather than *gestured, or gazed*)? Isn't it hard enough to get spoken language recognition, reasoning, a discourse model, and all the rest of the essential components of dialogue working without having to worry about non-verbal behaviors (which, the skeptic might say, aren't even important in human-human dialogue)?

There are three reasons why it might and should occur to us to add the non-verbal modalities one comes purely from the human side of things, the second and third come from the interaction between computer and human. First, it occurs to us to add the non-verbal modalities to dialogue systems as soon as we take a close look at what really goes on in human-human dialogue. To be sure, we can speak on the telephone with one another and make ourselves understood perfectly well but, when we are face-to-face with another human, no matter what our language, cultural background, or age, we all use our faces and hands as an integral part of our dialogue

---

This research was funded by the NSF STIMULATE program, Deutsche Telekom, and the other generous sponsors of the MIT Media Lab.

with others. Second, we may turn to the non-verbal modalities when we reflect on the difficulties we have getting users to behave as they need to when interacting with perfectly adequate spoken dialogue systems. Users repeat themselves needlessly, mistake when it is their turn to speak, and otherwise behave in ways that make dialogue systems *less* likely to function well [20]. It is in situations just like these in life that the non-verbal modalities come in to play: in noisy situations, humans depend on access to more than one modality [26]. This leads us to the third reason we might wish to add the non-verbal modalities to dialogue systems. While humans have long years of practicing communication with other humans (some might even say this ability is innate [31]), communication with machines is learned. And yet, it has been shown that given the slightest chance, humans will attribute social responses, behaviors, and internal states to computers [24].

If we can skillfully build on that social response to computers, channel it even into the kind of response that we give one another in human conversation, and build a system that gives back the response (verbal and nonverbal) that humans give, then we may evoke in humans the kinds of communicative dialogue behaviors they use with other humans, and thus allow them to use the computer with the same kind of efficiency and smoothness that characterizes their human dialogues. There is good reason to think that non-verbal behavior will play an important role in evoking these social communicative attributions. Our research shows that humans are more likely to consider computers life-like human-like even when those computers display not only speech but appropriate nonverbal communicative behavior.

What non-verbal behaviors, then, do humans fruitfully use with other humans to facilitate dialogue? Spontaneous (that is, unplanned, unselfconscious) gesture accompanies speech in most communicative situations, and in most cultures (despite the common belief to the contrary, in Great Britain, for example). People even gesture while they are speaking on the telephone [25]. We know that listeners attend to such gestures in face-to-face conversation, and that they use gesture in these situations to form a mental representation of the communicative intent of the speaker [6]. Likewise, faces change expressions continuously, and many of these changes are synchronized to what is going on in concurrent conversation. Facial displays are linked to the content of speech (winking when teasing somebody), emotion (wrinkling one’s eyebrows with worry), personality (pouting all the time), and other behavioral variables. Facial displays can replace sequences of words (she was dressed [wrinkle nose, stick out tongue]) as well as accompany them [10],and they can serve to help disambiguate what is being said when the acoustic signal is degraded. They do not occur randomly but rather are synchronized to one’s own speech, or to the speech of others [8]; [13]. Eye gaze is also an important feature of non-verbal communicative behaviors. Its main functions are to help regulate the flow of conversation; that is, to signal the search for feedback during an interaction (gazing at the other person to see whether s/he follows), to signal the search for information (looking upward as one searches for a particular word), to express emotion (looking downward in case of sadness), or to influence another person’s behavior (staring at a person to show that one won’t back down) [3], [9].

Although there are many kinds of gestures and an almost infinite variety of facial displays [1], the computer science community for the most part has only attempted to integrate *emblematic* gestures (e.g. the thumbs up gesture, or putting one's palm out to mean stop), that are employed in the absence of speech, and *emotional* facial displays (e.g. smiles, frowns, looks of puzzlement) into the construction of human-computer interface system. But in building dialogue systems we want to exploit the power of gestures that function in conjunction with speech. And emotions are inappropriate in the majority of situations for which we use automatic dialogue systems. We would not expect to have a weather system be *sad*, even if it's raining in New York. Most importantly, the regulative functions of both kinds of non-verbal behaviors (e.g. to facilitate smooth turn-taking, or give feedback) have been ignored, and it is these functions that promise to improve the performance of spoken dialogue systems.

For the construction of dialogue systems, then, there are types of gestures and facial displays that can serve key roles. In natural human communication, both facial displays and gesture add redundancy when the speech situation is noisy, both facial displays and gesture give the listener cues about where in the conversation one is, and both facial display and gesture add information that is not conveyed by accompanying speech. For these reasons, facial display, gesture and speech can profitably work together in *embodied dialogue systems.* In this chapter, I argue that the functions of non-verbal behaviors that are most valuable to spoken dialogue systems are those that are finely timed to speech, integrated with the underlying structure of discourse, and responsible for the regulation of conversation. These are the reasons to *embody* spoken dialogue systems.

Two Examples

Let's look at how humans use their hands and faces. In the following picture, Mike Hawley, one of my colleagues at the Media Lab, is shown giving a speech about the possibilities for communication among objects in the world. He is known to be a dynamic speaker, and we can trace that judgment to his animated facial displays and quick staccato gestures. As is his wont, in the picture below Mike's hands are in motion, and his face is lively. As is also his wont, Mike has no memory of having used his hands when giving this talk. For our purposes, it is important to note that Mike's hands are forming a square as he speaks of the mosaic tiles he is proposing to build. His mouth is open and smiling and his eyebrows raise as he utters the stressed word in the current utterance. Mike's interlocutors are no more likely to remember his non-verbal behavior than he. But they do register those behaviors at some level, and use them to form an opinion about what he said, as we will see below.

Figure 1: Talking about mosaic tiles

Figure 2 shows Seymour Papert, another colleague of mine, talking about embedding computing in everyday objects and toys. He breathes in, looks up to the right, then turns towards the audience and says "A kid can make a device that will have real behavior (...) that two of them [will interact] in a – to – to do a [dance together]". When he says "make a device" he looks upward; at "real behavior" he smiles; on "will interact" he looks towards the audience, raises his hands to chest level and points with each hand towards the other as if the hands are devices that are about to interact. He holds that pointing position through the speech disfluency and then, while saying "dance together", his hands move towards one another and then away, as if his fingers are doing the tango (not shown). He then looks down at his hands, looks to the side and pauses, before going on.

Figure 2: ". . . will interact"

Now, because this is a speech, and not a conversation, some of the integration with verbal and nonverbal behavior is different than in conversation, but some is also strikingly the same. For example, although nobody else is taking a turn, Papert still gazes away before taking the turn, and gazes towards his audience as he begins to speak. Likewise, he gazes away at the end of this particular unit of the discourse, and then turns back as he continues. Papert also uses the four kinds of gestures that are found in conversations. His gestures are still aligned with the most prominent phonological units of his speech and, there is co-articulation such that the first gesture, a deictic (pointing gesture), perseverates through his speech disfluency, allowing the second gesture, an iconic (representational gesture), to co-occur with the semantic unit it most resembles.

Such a performance is repeated almost anytime anybody speaks with anybody else. And, as I will discuss further below, when these nonverbal cues are missing, people become more disfluent, and less able to achieve smooth turn-taking. It is for exactly these reasons that we incorporate these non-verbal behaviors in spoken dialogue systems, creating embodied conversational agents capable of exploiting multiple channels for conveying information, and also for regulating the conversation.

A Demonstration

Before we turn to a description of how our embodied conversational agents work, let's start with an example of their behavior. In the system that I will describe here one embodied agent talks to another. While there is no human participant in these dialogues, the example is closer to what we would like to achieve for human-system dialogue (in the absence of engineering problems, such as perfect speech and gesture recognition), and so will serve to illustrate the non-verbal behaviors that we would like to integrate into automatically generated spoken language.

For this example, imagine that Gilbert is a bank teller, and George, a customer, has asked Gilbert for help in obtaining $50 (as the dialogue is generated automatically the two agents have to specify in advance each of the goals they are working towards and steps they are following; this explains the redundancy of the dialogue).

Gilbert: Do you have a blank check?

George: Yes, I have a blank check.

Gilbert: Do you have an account for the check?

George: Yes, I have an account for the check.

Gilbert: Does the account contain at least $50?

George: Yes, the account contains $80

Gilbert: Get the check made out to you for $50 and then I can withdraw $50 for you.

George: All right, let's get the check made out to me for $50.

In this example, as in (American) life, the yes/no questions end on rising intonational contours, and the answers end on falling intonational contours. The most accented syllable is determined by which information is new and salient. Information about which words or phrases are most salient to the discourse, whether words or phrases refer to places in space, or spatializable entities, and which words or phrases end a speaker's turn also determine the placement and content of gestures and facial displays.

In particular, when Gilbert asks a question, his voice rises. When George replies to a question, his voice falls. When Gilbert asks George whether he has an account for the check, he stresses the word "account". When he asks whether George has a blank check, he stresses the word "check". Every time Gilbert replies affirmatively ("yes"), or turns the floor over to George (at the ends of utterances), he nods his head, and raises his eyebrows. George and Gilbert look at each other when Gilbert asks a question, but at the end of each question, Gilbert looks up slightly. During the brief pause at the end of affirmative statements the speaker (always George, in this fragment) blinks. To mark the end of the questions, Gilbert raises his eyebrows. In saying the word "account", Gilbert forms a kind of box in front of him with his hands: a metaphorical representation of a bank account in which one keeps money. In saying “check”, Gilbert sketches the outlines of a checkbook in the air between him and his listener.

In Figure 3 and Figure 4 are reproduced excerpts from the conversation.

Figure 3: (a) "do you have a blank check?"; (b) “can you help me?”

Figure 4: (a) "you can write the check"; (b) “I have eighty dollars”

Figure 3(a) shows the automatic generation of an iconic gesture representing a check or checkbook, along with the phrase “do you have a blank check”; (b) shows the generation of a metaphoric gesture representing supplication, along with the phrase “can you help me”.

Figure 4(a) shows the automatic generation of an iconic gesture indicating writing on something, along with the phrase “you can write the check”, and (b) shows the generation of a beat gesture along with the phrase “yes, I have eighty dollars in my account”.

Embodied Dialogue Systems

In moving from studying conversation between humans, such as that exemplified by Figures 1 and 2, to implementing computer conversations, such as that illustrated in Figures 3 and 4, we are moving from a rich description of a naturally occurring phenomenon to a parametric implementation. In the process, certain aspects of the phenomenon emerge as feasible to implement, and certain aspects of the phenomenon emerge as key functions without which the implementation would make no sense. In this section we address these two issues: what *can* we do when computers talk to humans (or to other computers, as in Figures 3 and 4), and what cannot we not afford to leave out, if we believe that nonverbal behavior is of any utility to automatic dialogue systems.

Three testbed projects address the aspects of non-verbal behavior that we can implement and must implement when we build embodied dialogue systems.

Animated Conversation

In the first embodied conversational agent that I will discuss (created in conjunction with Norm Badler, Mark Steedman, Catherine Pelachaud, and students at the University of Pennsylvania's Human Simulation Laboratory), the goal was to derive multimodal (speech with intonation, facial displays and hand gestures) output from one single representation of propositional content. If gestures and facial displays are tightly coupled to the underlying discourse structure of speech, and to intonation in production (so the argument went), then we should be able to generate all of these behaviors as part of one single process in embodied dialogue. At that point in time we were not ready to address the problems of *understanding* human multimodal behavior, and so we built two embodied dialogue systems that could converse with one another, using Badler's work on human figure animation [1] as the body. Our focus in the Animated Conversation system was the choice of the right non-verbal behavior to generate (which kind of facial display, and which of the four types of gesture), and then the alignment of that non-verbal behavior to the verbal behavior with respect to the temporal, semantic, and discourse aspects of the dialogue.

The Dialogue Planner

At the top of this system is a dialogue generation engine inspired by Power [22], but enriched with explicit representations of the structure of the discourse and the relationship of the structure to the agents' domain plans. These added representations, which describe the entities that are of discourse concern and the purposes for which agents undertake communicative actions, figure crucially in the determination of the appropriate gesture and intonation to accompany agents' utterances.

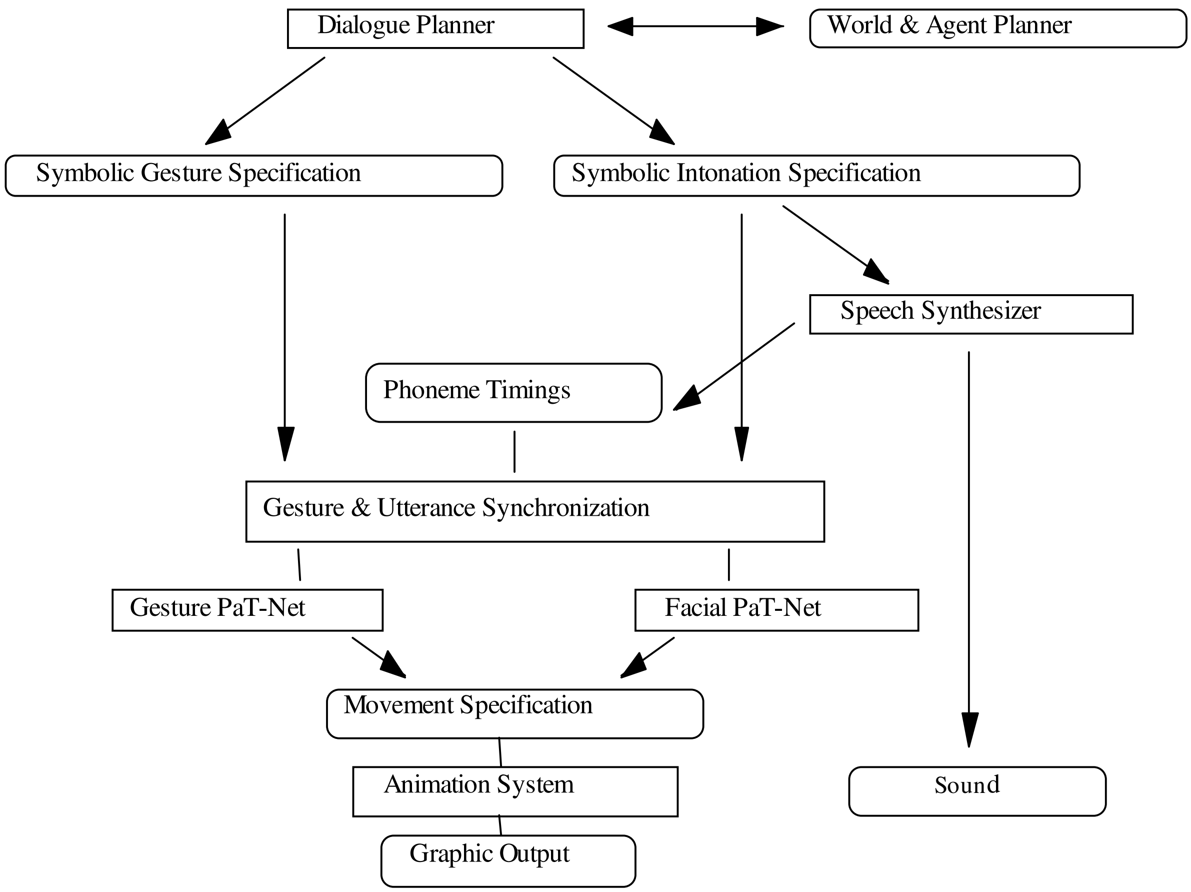

Figure 5: Architecture of Animated Conversation

The input to this engine is a database of facts describing the way the world works, the goals of the agents, and the beliefs of the agents about the world, including the beliefs of the agents about each other. This specification may assign not only different goals to the agents, but different beliefs and different capabilities for action as well. Each such distinction may potentially influence the course of the dialogue. In our example, the customer's goal of obtaining $50 motivates the dialogue; the customer's ability to write his check and the teller's ability to complete the transaction determine how the two settle on a plan; and the customer's readiness to write a check settles the conclusion of the dialogue.

The engine transforms this abstract input into a dialogue by running a simple hierarchical planner for each agent, using that agent's goals and beliefs. In this planner, certain kinds of goal expansions and action executions trigger instructions to take communicative actions. When such an instruction is generated, an agent suspends planning, and constructs a linguistic output (with gesture and intonation) corresponding to the instruction. The output is sent to the other agent, who computes on its own and ultimately returns its next contribution to their discourse. Depending on this message received in response, the agent may have to modify its plans or engage in further communication before reinvoking the planner. For instance, our dialogue illustrates two kinds of plan revision: upon hearing the customer's request to get

$50, the teller must add that as a new goal to his plans; and after the teller's proposal of a subplan, the customer must expand his goal of obtaining $50 into the steps the teller proposes. Examples of utterances in our dialogue generated in response to others include indications of agreement, indications of acknowledgment, and the answers to questions.

In performing revisions and replies, each agent must rely on additional knowledge and on established coordination between the agents, in order to determine when and how discourse actions are to be carried to completion. In particular, constructing a response requires knowledge about how communicative actions relate to one another, while modifying plans correctly requires understanding the significance of the response received, and maintaining links between parts of the plan and the discourse actions which may necessitate the revision of those parts. The agents are assumed to share their knowledge about dealing with discourse actions, as they share the knowledge in the planner that determines when discourse actions are appropriately initiated. At any point, then, each agent has interlinked representations of the domain plan that is being executed, and of the constituents of discourse that go into the discussion of the plan. In addition, a model of attention (the attentional state) indicates which entities are known to the participants, which entities have been referred to in the discourse, and how salient those entities are. After the teller asks whether the customer has a blank check, for example, the customer and check are listed in the attentional state as most salient, while the teller, the account and the $50 it contains are less salient. In addition, a record of the purposes generated by the planner that initiated discourse actions is kept.

The most important use of the explicit attentional and intentional state of the discourse is in annotating the logical representations produced by the dialogue generator for the pragmatic factors that determine what intonation contours and gestures are appropriate in its linguistic realization. For intonation, each node in the logical representation is labeled according to the status of the information it presents in the discourse: whether it is part of the theme or the rheme. Material is classified as thematic if it occurs in part of the speaker's discourse purpose in the current constituent or its ancestors for which evidence has been given. Meanwhile, material is classified as part of the rheme if it occurs only in that part of the speaker's discourse purpose in the current segment or its ancestors for which textual evidence has not yet been provided. Given this annotation, text is generated and pitch accents and phrasal melodies are placed on generated text roughly as outlined in Steedman [27] and Prevost and Steedman [23]. In declarative sentences, rhematic information gets pitch accents wherever possible, and is presented with a rise-fall intonation. In contrast, thematic information in declaratives is given a pitch accent and a distinct intonational contour only if contrastive (that is, only if referring to an entity when another would be more salient in that context), and receives a rise-fall-rise intonational contour. Unimportant information is never accented or assigned a separate intonational contour. The result is English text annotated with intonational cues. This text is converted automatically to a form suitable for input to the AT&T Bell Laboratories’ TTS synthesizer [16]. The resulting speech and timing information is then critical for synchronizing the facial and gestural animation.

Symbolic Gesture Specification

This discourse and intonation infrastructure allows us to generate types of gestures, and placement of gestures as follows. Utterances are annotated according to how their semantic content could relate to a spatial expression (literally, metaphorically, spatializably, or not at all). These annotations result in the association of gesture to content in the following way:

- Concepts that referred to entities with a physical existence in the world were accorded iconics/representational gestures(concepts such as 'checkbook', 'write', etc.).
- Concepts with common metaphoric vehicles received metaphorical gestures (concepts such as 'withdraw [money]', 'bank account', 'needing help');
- Concepts referring to places in space received deictic/pointing ('here', 'there').
- Formless baton-like beat gestures were generated for items where the semantic content cannot be represented, but the items were still unknown, or *new*, to the hearer (the concept of "at least").

If a representational gesture is called for, the system accesses a dictionary of gestures for concepts in order to determine the symbolic representation of the particular gesture to be performed.[6]

After this gestural annotation of all gesture types, and lexicon look-up of appropriate forms for representational gestures, information about the duration of intonational phrases (acquired in speech generation) is used to time gestures. First, all the gestures in each intonational phrase are collected. Because of the relationship between accenting and gesturing, in this dialogue, at most one representational gesture occurs in each intonational phrase. If there is a representational gesture, its preparation is set to begin at or before the beginning of the intonational phrase, and to finish at or before the next gesture in the intonational phrase, or the nuclear stress of the phrase, whichever comes first. The stroke phase is then set to coincide with the nuclear stress of the phrase. Finally, the relaxation is set to begin on the end of the stroke or the end of the last beat in the intonational phrase, with the end of relaxation to occur around the end of the intonational phrase. Beats, in contrast, are simply timed so as to coincide with the stressed syllable of the word that realizes the associated concept. When these timing rules have applied to each of the intonational phrases in the utterance, the output is a series of symbolic gesture types and the times at which they should be performed. These instructions are used to generate motion files that run the animation system.

Symbolic Facial Specification

Facial displays were generated both as a function of the dialogue structure and turn-taking structure. A character was more likely to look at the other if his utterance was particularly short, if the utterance was a question, if he was accenting a word, and at the end of his turn to speak. He was more likely to look away if he was about to produce an utterance, if he was answering a question or carrying out a request, or if he was signaling to the other that he would not take the turn during the other's within-turn pause. Characters nodded when they were acquiescing (semantic nods), and produced short nods along with pitch peaks on lexical items and as feedback signals (during a grammatical pause on the part of the other character during the other character's turn)[7].

Lessons Learned

The literature on the association of verbal and non-verbal behavior has for a very long time been purely descriptive. The goal of Animated Conversation was to see if we could parameterize those descriptions, and make them into predictive rules. Our evaluation took the form of showing the animation to various lay people and experts in non-verbal behavior and asking them what looked right and what looked wrong. Two important issues were brought out in this way. First, we realized that while a discourse framework could specify type of gesture and placement of gesture, we would need a semantic framework to generate the *form* of particular gestures. In the Animated Conversation system we were obliged to choose gestural forms from a dictionary of gestures. That was a hack that we were uncomfortable with. We didn't generate the *form* of the gestures from scratch, and so although we took advantage of what we knew in terms of temporal integration and discourse integration, we didn't exploit rules for semantic integration. Likewise, we realized in watching the animation that *too many* nonverbal behaviors were being generated—the impression was of a bank teller talking to a foreigner, and trying to enhance his speech with supplementary nonverbal cues. This problem arose because each nonverbal behavior was generated independently, on the basis of its association with discourse and turn-taking structure and timed by intonation, but without reference to the other nonverbal phenomena present in the same clause. Our conclusion was that we lacked two functions in our system: first, a multimodal "manager" that distributes meaning across the modalities, but that is essentially modality-independent in its functioning. Such "managers" have been described for multimodal integration for generation of text and graphics [32], and multimodal integration in input [12]. Second, we lacked an understanding of what shape a particular gesture would take: how did we describe which particular gesture would be generated? This is similar to the problem of word choice in text generation (Elhadad et al. to appear). We will return to our current approach to these difficult issues below.

Gandalf

Animated Conversation was designed to generate non-verbal behaviors as a function of the underlying propositional content of a dialogue. Some non-verbal behaviors, however, are not predictable from propositional structure and are rather determined by the *interactional structure* of a conversation. Gaze behavior, for example, is predictable in part from the information structure of a dialogue, and in part predictable from the turn-taking structure of the conversation. As described above, for example, we look at each other when we give over the turn. Gandalf is a system designed by Kristin Thorisson in conjunction with others in my research group to generate—and to understand—non-verbal behaviors with an interactional function. This meant that many of the same behaviors as were generated by Animated Conversation, were generated by Gandalf, but as a function of conversational interaction, rather than discourse structure. And whereas the conversation in Animated Conversation involved two autonomous agents, Gandalf can sustain a conversation with a human user, making these interactional behaviors especially important.

Gandalf was built within Ymir, a testbed system especially designed for prototyping multimodal agents that interpret human communicative behavior, and generate integrated spontaneous verbal and nonverbal behavior of their own (see [29], [30] for more details about the system). Ymir is constructed as a layered system. It provides a foundation for accommodating any number of interpretive processes, running in parallel, working in concert to interpret and respond to the user's behavior. In this way, Ymir offers opportunities to experiment with various computational schemes for handling specific subtasks of multimodal interaction, such as natural language parsing, natural language generation and selection of multimodal acts.

Ymir’s strength is the ability to accommodate two types of behavior described above, *communication* and *propositional*. As described above, some communicative behavior controls the *envelope of communication* . For example, gaze is an indicator to the participants of a conversation of who should speak when: when the current speaker looks at the listener and pauses, this serves as a signal that the speaker is giving up the turn [9]. On the other hand, some communicative behavior controls the *propositional content of communication*. For example, the content of speech, and the content of iconic gestures determine the direction that the conversation is taking. The envelope behaviors can be referred to as *reactive*, in that they are not reflected upon, nor do they convey particular content. In this they can be contrasted with the contentful reflective behaviors. Ymir has layers dedicated to reactive behaviors such as gaze and other turn-taking signals, reflective behaviors such as speech and contentful gestures, and process control. Reactive behaviors require fast "automatic" reactions to maintain the conversation (when the other interlocutor stops speaking and looks at me, I should begin to speak). This reactive layer in Ymir is differentiated from the reflective layer, which attends to speech input, the content of gestures, and other types of information that will need to be understood and responded to. The process layer contains modules that can use the state of other modules as input. For example, the job of generating filler speech such as "right, umm, let's see" when the content layer is slow to generate speech or to finish speech processing, falls to the process control layer. The *action scheduler* takes commands from the other modules and negotiates the resources needed for each command to be obeyed. If a command is sent from the content layer asking for speech about a planet at the same time as the reactive layer sends a request for Gandalf to produce some kind of feedback acknowledgment, then the action scheduler may choose to generate the feedback acknowledgment as a non-verbal behavior (a nod, for example) so that the mouth is free to produce content-oriented speech.

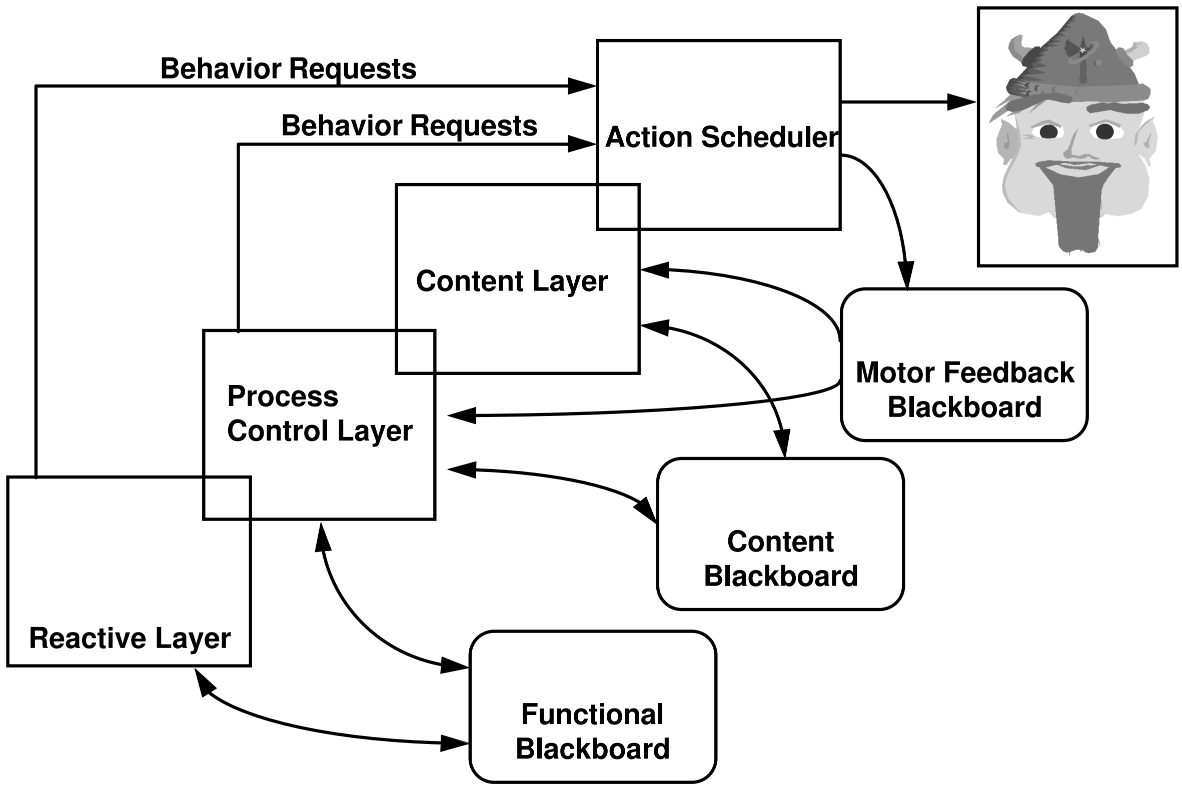

Figure 6: Gandalf, blackboard architecture

Gandalf is the first agent constructed in the Ymir architecture. It has been provided with the minimal behaviors necessary for face-to-face dialogue. It ‘understands’ body stance (oriented towards Gandalf or towards the task at hand)[8], and the function of some hand gestures. It understands the social conventions of gaze and head/face direction and integrates those to provide the correct feedback behaviors at the correct time. In particular, Gandalf uses eye gaze to regulate turn-taking, nods to signal that he is following the user’s speech, and beat gestures to take the turn, and to indicate that he is answering a question. Gandalf understands pointing gestures, gaze as an indication of turn-taking, and body orientation as an indication of conversation- or task-oriented activity. The prototype primarily serves to demonstrate Ymir's treatment of the timing of multimodal acts, and to illustrate Ymir's ability to accept and integrate data from independent modules that work on partial input data, and to integrate data at multiple levels. The Gandalf system does not generate speech but rather chooses from some pre-canned utterances.

People interact with Gandalf by putting on a jacket and thin gloves which allow the system to sense the position of the body with respect to the screen showing Gandalf’s face and the screen showing a map of the solar system. Gandalf is introduced as an expert on the solar system. Users can ask to be shown particular planets, and can ask for information about those planets. Gandalf interprets gestures pointing towards the screen as references to planets, and also interprets turns towards the screen as initiations of task activity.

Lessons Learned

Gandalf is a successful first implementation of an embodied dialogue agent. Unlike Animated Conversation, it is able to converse with people and to interpret some non-verbal behavior in synchrony with spoken language. The many people, both adults and children, who have interacted with Gandalf over the last two years have found the interaction satisfying, engaging, and natural. It is notable that people interacting with the system begin by standing still in front of the screen with the solar system and, within two conversational turns, begin to adopt much more spontaneous and human-conversational movements. That is, they begin to look at Gandalf when it is speaking, but look at the solar system when Gandalf is showing them a planet (note that there is no objective *need* for them to look at Gandalf's face. All of the propositional content is displayed on the solar system screen, or conveyed via spoken language). They begin to nod when Gandalf is speaking, as if to give feedback, and so forth.

On the other hand, Gandalf is lacking a body (only Gandalf's face and a single hand are shown) and so the range of hand gestures available to Animated Conversation are lacking here. In addition, users quickly run through the canned utterances that Gandalf can produce—generation of language is clearly needed. An evaluation of Gandalf (see below), convinces us that envelope feedback behaviors are important to embodied spoken dialogue systems. But, Animated Conversation convinced us that gestures are also important, and those are lacking here.

Our most recent project attempts to carry one step further the possibilities of non-verbal behavior in spoken dialogue systems by integrating the interactional and propositional aspects of verbal and non-verbal behavior. In the next section I talk about the Rea system.

Real Estate Agent: Integration

While Animated Conversation, and Gandalf represent significant advances in autonomous, embodied, conversational agents, neither system is complete. The agents in Animated Conversation cannot interact with real people in real time. Gandalf, on the other hand, fails to model planning, language and propositional non-verbal behaviors at a level necessary for accomplishing non-trivial tasks. In order to overcome these deficiencies, the next generation of animated, conversational characters must integrate the propositional and interactional layers of communication, and account for their interactions. One of the difficulties in reaching this goal, however, is that the constraint of running in real time require a trade off between linguistic processing and generation, and reactivity. Our goal, then, was to design Rea in such a way that its reactions are aptly timed, and might even provide more time for the reflective layer to come up with the correct content, either in interpretation or generation. That is, a well placed "hmm, let's see" with slow thoughtful nods of the head can give users necessary information about the state of the conversation, while allowing the system a little more time to come up with a contentful answer. In other words, Rea was constructed from a suite of communication skill processes which are operating with different response times.

Figure 7: Rea, the Realtor

The domain we chose was real estate: Rea can interact with a human around the purchase of a home. We chose this domain for the importance that social interaction as well as knowledge plays in the success of a conversation. That is, real estate agents typically come to know the needs and desires of their clients through casual social interaction as well as check-lists. Part of our development efforts concentrates on making sure that Rea will be able to engage in social chit-chat, as well as being able to take users through 3D walk-throughs of different houses on a large screen.

A key area in which Rea has roots both in Animated Conversation and Gandalf is turn-taking. In "Animated Conversation," turns are allocated by a planner that has access to both agents' goals and intentions. In the real world, turns are negotiated through interactional non-verbal cues, rather than strictly imposed by the structure of the underlying task. In Gandalf, turns are negotiated by the agent and the human user in a natural way. Since Gandalf does no dialogue planning, however, each turn is artificially restricted to a single utterance. To competently handle turn-taking a system must interleave and process the propositional and interactional information in a principled way.

Some of the areas which we are exploring in the Rea architecture are the following:

- Use of verbal and non-verbal cues in interpretation. In a multimodal conversation system, the interpretation component must not only integrate information from different modalities into a coherent propositional representation of what the user is communicating, but—in order to know what function that information fills in the ongoing conversation—it must also derive the interactional information from the perceptual inputs. Moreover, it must determine when the user has communicated enough to begin analysis and be able to re-analyze input in case it misinterprets the user's turn-taking cues.

- The role of non-verbal behaviors in dialogue planning. The discourse planner for conversational characters must be able to plan turn-taking sequences and easily adapt when those plans are invalidated by non-verbal cues—for example when the human refuses to give over the turn, and continued nonverbal feedback becomes more appropriate than adding new content to the conversation.
- Generation of verbal and non-verbal behaviors. When the discourse plan calls for the generation of interactional information, the character must decide which modality to use, and must take into account interactional information from the user. For example, signaling an interruption or termination may be performed verbally or with a nod of the head depending on whether the user or the character currently has the turn. Crucially, Rea's architecture begins to fill our goal of *function-oriented* rather than modality-oriented processes. That is, rather than specifying what a gesture will

do at any given moment, the system generates a need for a particular function to be filled, and that modality that is free at that moment (and that the system knows capable of filling that function) is called into play.

Architecture

Figure 8 shows the modules of the Rea architecture. The three key aspects for Embodied Conversational Agents are:

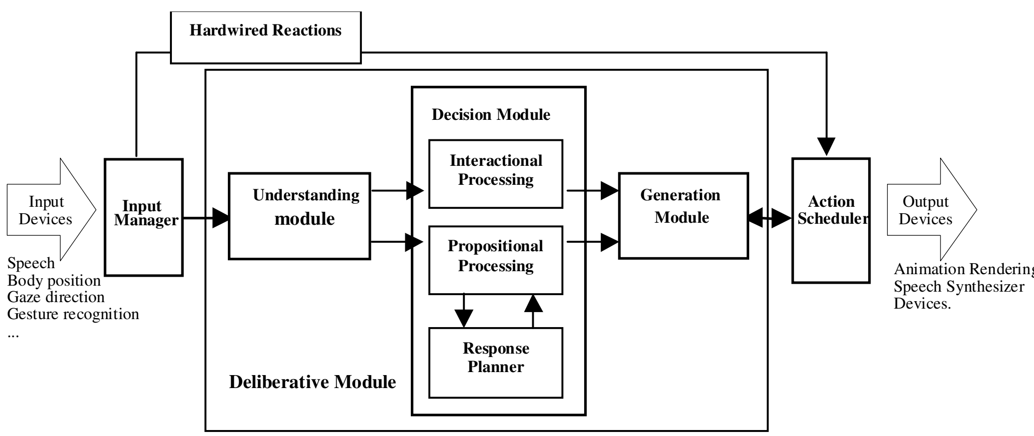

Figure 8: Rea Architecture

- 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 semantic 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.
- The categorization of behaviors in terms of their conversational functions is mirrored by the

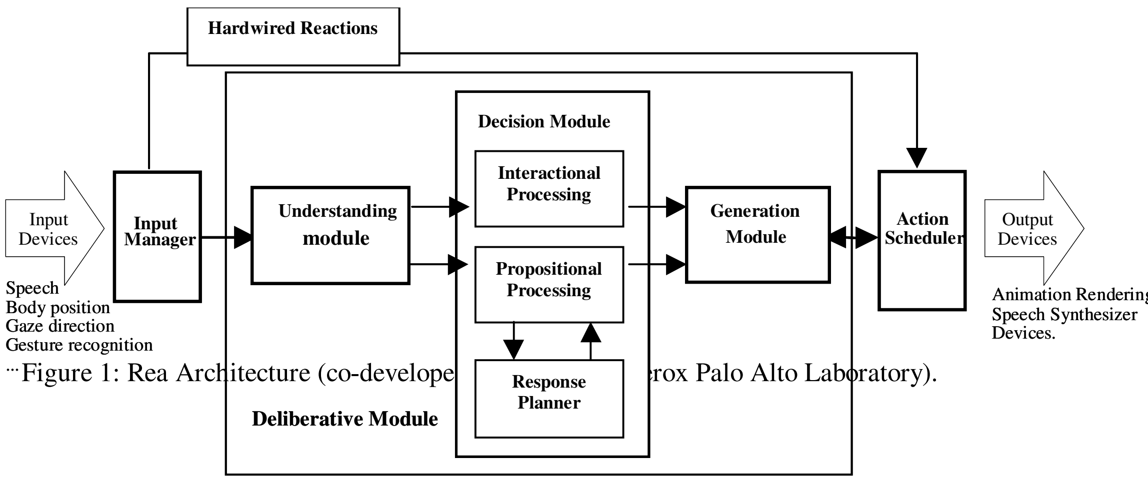

Figure 1: Rea Architecture (co-develope ... rox Palo Alto Laboratory).

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).

The Input Manager collects input from all modalities and decides whether the data requires instant reaction or deliberate discourse processing. Hardwired Reaction handles spontaneous reaction to stimuli such as the appearance of the user. These stimuli can then directly modify the agent's behavior without much delay. For example, the agent's gaze can seamlessly track the user's movement. The Deliberative Discourse Processing module handles all input that requires a discourse model for proper interpretation. This includes many of the interactional behaviors as well as all propositional behaviors. Lastly 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 prevent collisions between competing motor requests. The modules communicate with each other using KQML, a speech-act based inter-agent communication protocol, which serves to make the system modular and extensible.

Implementation

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 (written in SGI OpenGL) and conversation engine (written in C++ and CLIPS), while several other computers manage the speech recognition (until recently IBM Via Voice; currently moving to SUMMIT) and generation (previously Microsoft Whisper; currently moving to BT Festival) and image processing (STIVE).

In the implementation of Rea we have attended to both propositional and interactional components of the conversational model. In terms of the propositional component, Rea's speech and gesture output is generated in real-time. The descriptions of the houses that she shows, along with the gestures that she uses to describe those houses are generated using the SPUD natural language generation engine, modified so as to also generate natural gesture [7]. In this key aspect of Rea's implementation, speech and gesture are treated on a par, so that a gesture may be just as likely to be chosen to convey Rea's meaning as a word.

In the interactional component the following functions are possible:

- 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 "okay" in response to short pauses in the user's speech; she raises her eyebrows to indicate partial understanding of a phrase or sentence.
- Turn-taking 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 an expression of a desire to speak, and therefore halt her remarks at the nearest sentence boundary. Finally, at the end of her speaking turn she turns to face the user.

These conversational functions are realized 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 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. 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 1 summarizes Rea's current interactional output behaviors.

<<Table 1>>

Evaluation: Do Bodies Offer Anything to Dialogue Systems?

Because of Gandalf/Ymir's modular architecture, it is quite easy to build agents with different kinds of conversational skills. This flexibility allowed us to test two of the functions of non-verbal behavior described in the section on human-human dialogue. As I described above, one function of the face is to display envelope feedback, and another function is to display emotional expressions. A recent debate within the human-computer interface community centers on the importance of emotional expressions to human-like agents. In this literature, emotional feedback has meant *emotional emblems*, facial displays that reference a particular emotion without requiring the person showing the expression to feel that emotion at the moment of expression [10]. In the literature on anthropomorphism in interface systems, emotional feedback as displayed by the animated agent's emotional emblems in response to a user's input is held to be a feature that an embodied agent-based interface could—and *should*—add to human-computer interaction (cf.[11]; [14]; [18]; [28]). The emotional feedback used in such systems has been, in general, very simple: scrunched eyebrows to indicate puzzlement, a smile and raised eyebrows to indicate happiness. Thorisson and I claimed, on the contrary, that emotional emblems are not effective in conversational systems because they are not tightly integrated in function with the other behaviors generated. Our claim is that the importance of embodiment in computer interfaces lies first and foremost in its power as a *unifying concept for representing the processes and behaviors surrounding conversation*. If this is true, feedback that relates directly to the process of the conversation should be of utmost importance to both conversational participants, while any other variables, such as emotional displays, should be secondary. To test this hypothesis, we built three autonomous agents, all capable of full-duplex multimodal interaction (speech, intonation, and gesture in the input and output), but each giving a different kind of feedback (Cassell and Thorisson, forthcoming).

Agent #1 (Gandalf) gave content-related feedback only. That is, he was capable of executing commands and answering questions. An example of an interaction with an agent in the content condition follows:

<<Figure 9>>

Gandalf: "What can I do for you?" [*face looks at user. Eyes do not move.*]

User: "Will you show me what Mars looks like?" [*user looks at Gandalf.*]

Gandalf: "Why not—here is Mars" [*face maintains orientation. No change of expression. Mars appears on monitor.*]

User: "What do you know about Mars?" [*user looks at map of solar system.*]

Gandalf: "Mars has 2 moons" [*face maintains orientation. No change of expression.*]

Agent #2 (Roland) gave content feedback, but was also capable of emotional expressions. In particular, it produced a confused expression when it failed to interpret an utterance, and smiled when addressed by the user and when acquiescing to a request (for example to take the user to a particular planet). An example of an interaction with an agent in this emotional condition follows:

Gandalf: "What can I do for you?" [*Gandalf smiles when user's gaze falls on its face, then stops smiling and speaks*]

User: "Take me to Jupiter" [*user looks at screen and then back at Gandalf and so Gandalf smiles*]

Gandalf: "Sure thing. That's Jupiter" [*Gandalf smiles as it brings Jupiter into focus on the screen*]

User: [*Looks back at Gandalf. Short pause while deciding what to say to Gandalf.]*

Gandalf: [*looks puzzled because the user pauses longer than expected, waits for user to speak.*]

User: "Can you tell me about Jupiter?"

Agent #3 (Bilbo) gave content feedback, but was also capable of providing envelope feedback. In particular, this agent could turn its head and eyes towards the user when listening, and towards task when executing commands in the domain. It could avert its gaze and lift its eyebrows when taking turn. It gazed at the person when givingup its turn. Finally, it produced beat gestures when providing verbal content. An example of an interaction with this envelope agent follows:

User: “Is that planet Mars?”

Gandalf: “Yes, that’s Mars.” [*Gandalf raises eyebrows and performs beat gesture while saying “yes”, turns to planet and points at it while saying “that is Mars”, and then turns back to face user*]

User: I want to go back to Earth now. Take me to Earth [*user looks at map of solar system so Gandalf looks at solar system*]

Gandalf: “OK. Earth is third from the sun.” [*Gandalf turns to planet as it appears on the screen, then turns to user and speaks*]

User: "Tell me more." [*Gandalf takes about 2 seconds to parse the speech, but knows within 250 ms when the user gives the turn, so looks to the side to* *indicate taking over the turn, and the eyebrows go up and down during hesitations while parsing the user's utterance:*]

Gandalf "The Earth is 12,000 km in diameter" [*Gandalf looks back at the user and speaks.*]

We found that, as we expected, people preferred to interact with the agent capable of envelope feedback, and in their evaluations of the system rated the other two agents as no different from one another. In fact, one user, seated in front of the emotional agent, implored Gandalf "come on! Just let me know you're listening!". In addition, users' were more efficient with this agent, using fewer utterances to accomplish the same work. In more recent work, where users engage in a more collaborative task with Gandalf (the Desert Survival Task), we are obtaining similar results. Interestingly, we find that users rate the emotional agent as more friendly and warm, but rate the envelope agent as more helpful and more collaborative. Thus, if we can find contexts in which being warm is more important than being helpful, emotional agents will prevail; otherwise envelope behaviors, as predicted, facilitate the interaction, and are perceived as facilitatory by users.

It still remains to test the function of gestures in embodied dialogue systems. The research mentioned above [6] shows that users do take gesture into account when constructing a representation of the content of a monologue, and that the information that they received only in the gestural channel is just as likely to be re-narrated in speech. We also know that users attend to gestures in our dialogue systems. In the envelope condition described above, users often began to mirror Gandalf's gestures, ultimately producing beat gestures in parallel places to those chosen by Gandalf. We are just beginning to construct evaluation contexts for the use of propositional and interactional gestures using the Rea platform.

The results discussed in this section are much more optimistic about the role of non-verbal behaviors than those obtained using videoconferencing. For example, Whittaker and O'Conaill [33] tested whether video (videoconferencing) provided (a) cognitive cues that facilitate shared understanding; (b) process cues to support turn-taking, and (c) social cues and access to emotional information. Only the last kind of cue was found to be supported by video in communication. Key to their findings seems to be the fact that current implementations of video technology (even high quality video) have not been able to provide audio and video without significant time lags. This, of course, disrupts conversational process, giving the impression of providing vital non-verbal cues, but providing the cues in the wrong places. Embodied conversational systems like those presented here may be more likely to provide a testing ground for the role of these non-verbal behaviors, and a fruitful context for their use.

Next Steps

In talking about Animated Conversation, I mentioned that we were dissatisfied with the question of what form to generate for particular gestures, and how to negotiate which content is conveyed in which modality. A key aspect of our current development efforts concentrates on the issue of generating form from scratch in conjunction with natural language generation. This raises the issue of a representation language that is modality-free. I believe that a key component of a grammar that will be able to handle the issue of semantic form for gesture is a semantic representation scheme located at the sentence planning stage of generation. This scheme can encode the proper level of abstraction for concepts involving motion so that features such as manner, path, telicity, speed and aspect can be independently applied to the various modalities at hand. For example, the gesture module might generate the manner of a motion, while the spoken verb generates the path, or vice-versa. Thus, one might say "I went to the store" but produce a walking gesture with one's index and second finger. In this way, two semantic frames which each contain partial knowledge of the content to be generated are unified.

Related Work

Although the topic of 'believable animated agents' has recently received a fair amount of attention, resulting in a plethora of animated humanoid, animal, or fantasy actors, very few researchers have attempted to integrate their animated figures with the demands of spoken language dialogue. Ball et al. [2]'s work on the Persona project has similarities with our work. They are building an embodied conversational interface that will eventually integrate spoken language input, a conversational dialogue manager, reactive 3D animation, and recorded speech output. 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, however, 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). Because of the limitations of using a creature with wings and a beak, rather than hands and a face, all of the gestures and facial displays that they employ fall under the *emblematic* category, rather than those categories of non-verbal behaviors that are timed carefully with respect to speech, and which regulate the interaction.

Loyall and Bates [17] share our goal of real-time responsive language generation mixed with non-verbal behaviors (although they do not distinguish between non-verbal behaviors and other, non-communicative, behaviors such as looking at an object on the horizon and, to date, they have generated text rather than speech). However, the primary goal of the Oz group is to build believable 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. 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, and it is these latter functions that are most important to us. In addition, like Ball et al., the Oz group has chosen a non-human computer character—in this instance, about as far away from human as one can get since Loyall and Bates talk about language generation for Woggles, which look like marbles with eyes. Characters such as these can certainly evoke in us an awareness of emotional reactions, but their gross features disallow fine-grained timing or interaction of verbal and non-verbal behaviors and, quite obviously, their lack of hands precludes the use of gesture. 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.

Some researchers have attempted to create humanoid interactive systems. As described earlier in this chapter, several laboratories have created interactive systems represented by faces on a screen. However most of these efforts have (mistakenly, we believe) concentrated on displaying emotion to the exclusion of other functions of the face, and in the absence of language use. Nagao and Takeuchi [28], however, implemented a 'talking head' that understood human input and generated speech in conjunction with facial displays of very similar types to those described in this chapter. Despite their goal of using the face to regulate conversation, the generation of facial displays and speech in their system was not timed down to the phonological unit: facial displays were timed to whole utterances. Not surprisingly, therefore, while facial displays were found to be helpful to the interaction, the effect did not persevere, and was not stronger than a learning effect for the use of a text-only system. We argue that fine-grained timing is critical if one wishes to use the functionality of faces and hands to enhance and regulate dialogue.

Noma and Badler [19] have created a virtual human weatherman, based on the *Jack* human figure animation that was used for the Animated Conversation system presented in this chapter. 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. (forthcoming) and Beskow and McGlashan [4]. Lester et al. [15] have implemented a pedagogical agent that lives in a graphically rich virtual world simulating the routing system of the Internet. Their agent does produce deictic gestures (in conjunction with recorded human speech). This system is notable for having incorporated the insight that deictic gestures are more likely to occur in contexts of referential ambiguity.

Perlin and Goldberg [21] have created a scripting language and architecture for animated humanoid figures with the goal of allowing non-expert programmers to create interactive characters. The animated humanoid characters created using their IMPROV system have movements and postures that are strikingly realistic, based on their application of coherent noise functions to motion characteristics. As they incorporate the ability to use language into these systems, however, the use of noise functions rather than simulation models becomes problematic. For example, they give an example of a script that a character might follow called "No Soap, Radio" in which a joke is told. The script calls an external speech system to generate the speech, and then also calls the “Joke Gestures” script which chooses appropriate gestures based on the character’s personality. But, while gestures are certainly affected by personality, mood, and a number of other large scope phenomena, they are produced as a function of fine-grained interactional and discourse constraints. This is the key insight that we have derived from work on human-human conversation, and tested in our evaluation of embodied dialogue systems.

Conclusions

In sum, it appears that if one is going to go to the trouble of *embodying* (associating a body to) spoken dialogue systems, then one should exploit the conversational functions of the body, and one should be very careful to associate those functions to speech in a discourse- and interaction-sensitive manner. One cannot simply build the body on the one hand, and the dialogue system on the other, and then pair the two in output. As Brennan and Hulteen have pointed out [5], conversation is fundamentally collaborative, and dialogue systems can be improved by focusing on making sure that both interlocutors (the human and the computer) are given adaptive feedback and information about dialogue state. In human-human conversation these functions are often assumed by the body, and displayed through non-verbal behaviors. In building embodied conversational systems, then, the choice of what body parts to animate should come from the demands of dialogue (such as the need to regulate turn-taking), and the dialogue system should be built with both control of and input from the body model (such as the ability to generate gestures in conjunction with new information, and the ability to perseverate particular stretches of speech in order to synchronize with the production of gestures).

References

[1] Badler, N. I., C. Phillips, & B. L. Webber, *Simulating Humans Computer Graphics Animation and Control*. Oxford: Oxford University Press, 1993.

[2] Ball, G., D. Ling, D. Kurlander, D. Miller, D. Pugh, T. Skelly, A. Stankosky, D. Thiel, M. Van Dantzich, & T. Wax, "Lifelike Computer Characters: The Persona Project at Microsoft Research," in *Software Agents*, Bradshaw, J. M., Ed. Cambridge, MA: MIT Press, 1997.

[3] Beattie, G. W., "Sequential Temporal Patterns of Speech and Gaze in Dialogue," in *Nonverbal Communication, Interaction and Gesture: Selections from Semiotica*, Sebeok, T. A. & J. Umiker-Sebeok, Eds. The Hague: Mouton, 1981.

[4] Beskow, J. & S. McGlashan, "Olga - A Conversational Agent with Gestures," *Proceedings of Proceedings of the IJCAI '97 workshop on Animated Interface Agents - Making them Intelligent*, Nagoya, Japan, 1997.

[5] Brennan, S. E. & E. A. Hulteen, "Interaction and Feedback in a Spoken Language System: A

Theoretical Framework," *Knowledge Based Systems*, vol. 8, pp. 143-151, 1995.

[6] Cassell, J., D. McNeill, & K. E. McCullough, "Speech-Gesture Mismatches: Evidence for One Underlying Representation of Linguistic and Non-Linguistic Information," *Pragmatics and Cognition*, vol. 7, pp. 1-33, 1999.

[7] Cassell, J. & M. Stone, "Living Hand to Mouth: Theories of Speech and Gesture in Interactive Systems," *Proceedings of AAAI Fall Symposium: Psychological Models of Communication in Collaborative Systems*, Cape Cod, MA, 1999.

[8] Condon, W. S. & W. D. Osgton, "Speech and Body Motion Synchrony of the Speaker-Hearer," in *The Perception of Language*, Horton, D. H. & J. J. Jenkins, Eds.: Academic Press, 1971, pp. 150-184.

[9] Duncan, S., "Some Signals and Rules for Taking Speaking Turns in Conversation," in *Nonverbal Communication*, Weitz, S., Ed.: Oxford University Press, 1974.

[10] Ekman, P., "About Brows: Emotional and Conversational Signals," in *Human Ethology: Claims

and Limits of a New Discipline, von Cranach, M., K. Fopps, W. Lepenies, et al., Eds. New York: Cambridge University Press, 1979, pp. 169-249.

[11] Elliott, C., "I Picked up Catapia and Other Stories: a Multimodal Approach to Expressivity for 'Emotionally Intelligent' Agents," *Proceedings of First International Conference on Autonomous Agents*, Marina del Ray, CA, 1997.

[12] Johnston, M., P. R. Cohen, D. McGee, J. Pittman, S. L. Oviat, & I. Smith, "Unification-Based Multimodal Integration.," *Proceedings of Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics (ACL-97/EACL-97)*, Madrid, Spain, 1997.

[13] Kendon, A., "Some Relationships between Body Motion and Speech," in *Studies in Dyadic Communication*, Siegman, A. W. & B. Pope, Eds. Elmsford, NY: Pergamon Press, 1972.

[14] Koda, T. & P. Maes, "Agents with Faces: The Effects of Personification of Agents," *Proceedings of Presented at Human-Computer Interaction '96*, London, UK, 1996.

[15] Lester, J., S. Towns, C. Calloway, & P. FitzGerald, "Deictic and Emotive Communication in Animated Pedagogical Agents," in *Embodied Conversational Agents*, Cassell, J., J. Sullivan, S. Prevost, et al., Eds. Boston: MIT Press, 2000.

[16] Liberman, M. & J. Pierrehumbert, "Intonational Variance under Changes in Pitch Range and Length," in *Language Sound Structure: Studies in Phonology Presented to Morris Halle by His Teacher and Students*, Aronoff, M. & R. T. Oehrle, Eds. Cambridge, MA: MIT Press, 1984, pp. 157-233.

[17] Loyall, A. & J. Bates, "Personality-Rich Believable Agents that Use Language.," *Proceedings of Agents '97*, Marina del Rey, CA, 1997.

[18] Nagao, K. & A. Takeuchi, "Social Interaction: Multimodal Conversation with Social Agents," *Proceedings of 12th National Conference on Artificial Intelligence (AAAI-94)*, Seattle, WA, 1994.

[19] Noma, T. & N. Badler, "A Virtual Human Presenter," *Proceedings of Workshop on Animated Interface*

Agents - Making them Intelligent (IJCAI'97), Nagoya, Japan, 1997.

[20] Oviatt, S. L., "Predicting spoken disfluencies during human-computer interaction," *Computer Speech and Language*, vol. 9, pp. 19-35, 1995.

[21] Perlin, K. & A. Goldberg, "Improv: A System for Scripting Interactive Actors in Virtual Worlds," *Proceedings of SIGGRAPH 96*, New Orleans, LA, 1996.

[22] Power, R., "The Organisation of Purposeful Dialogues," *Linguistics*, vol. 17, pp. 107-152, 1977.

[23] Prevost, S. & M. Steedman, "Specifying intonation from context for speech synthesis.," *Speech Communication*, vol. 15, pp. 139-153, 1994.

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

[25] Rime, B., "The Elimination of Visible Behavior from Social Interactions: Effects of Verbal, Noverbal and Interpersonal Variables," *European*

Journal of Social Psychology, vol. 12, pp. 113-129, 1982.

[26] Rogers, W. T., "The contribution of kinesic illustrators toward the comprehension of verbal behavior within utterances," *Human Communicatin Research*, vol. 5, pp. 54-62, 1978.

[27] Steedman, M., "Structure and intonation," *Language*, vol. 67, pp. 190-296, 1991.

[28] Takeuchi, A. & K. Nagao, "Communicative Facial Displays as a New Conversational Modality," *Proceedings of InterCHI 93*, Amsterdam, Netherlands, 1993.

[29] Thorisson, K. R., "Face-to-Face Communicatin with Computer Agents," *Proceedings of AAAI Spring Symposium on Believable Agents Working Notes*, Stanford University, CA, 1994.

[30] Thorisson, K. R., "A Mind Model of Multimodal Communicative Creatures and Humanoids," *Applied Artificial Intelligence*, vol. 13, 1999.

[31] Trevarthen, C., "Sharing Makes Sense: Intersubjectivity and the Making of an Infant's

Meaning," in *Language Topics: Essays in Honour of M. Halliday*, vol. 1, Steele, R. & T. Threadgold, Eds. Amsterdam: J. Benjamins, 1986, pp. 177–200.

[32] Wahlster, W., E. André, W. Graf, & T. Rist, "Designing Illustrated Texts," *Proceedings of 5th EACL*, Berlin, Germany, 1991.

[33] Whittaker, S. & B. O'Conaill, "The Role of Vision in Face-to-Face and Mediated Communication," in *Video-Mediated Communication*, Finn, K. E., A. J. Sellen, &S. B. Wilbur, Eds. Hillsdale, NJ: Lawrence Erlbaum Associates, 1997, pp. 23–49.