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.

Because most researchers in the field of spoken dialogue may not be familiar with the literature on how speech and non-verbal behaviors are integrated in humans, the form of this chapter will be as follows: first I will describe how gesture and facial/head movements are used by humans and how these non-verbal behaviors are integrated into conversation among humans. Next I will turn to some of the issues that arise when we attempt to use the information available to us about natural human communication in the design of embodied dialogue systems. I will describe these issues in the context of a number of prototype systems that my students and I have built relying on these principles. Finally, I will discuss the evaluation of these systems, and whether non-verbal behaviors add anything to dialogue systems.

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

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 with others. I will return to this point at length below, but for the moment, take the statement on faith. 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 (Oviatt, 1995). 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 (Rogers, 1978). 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; Trevarthen, 1986), 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 (Reeves & Nass, 1996).

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 (Rimé, 1982). 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 (Cassell et al, 1998). 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 (Ekman, 1979), 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 (Condon & Osgton, 1971; Kendon, 1972). 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) (Beattie, 1981; Duncan, 1974).

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.

[1] Following Takeuchi & Nagao 1993, we use the term "facial display" rather than "facial expression" to avoid the automatic connotation of emotion that is linked to the latter term.

An Example

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.

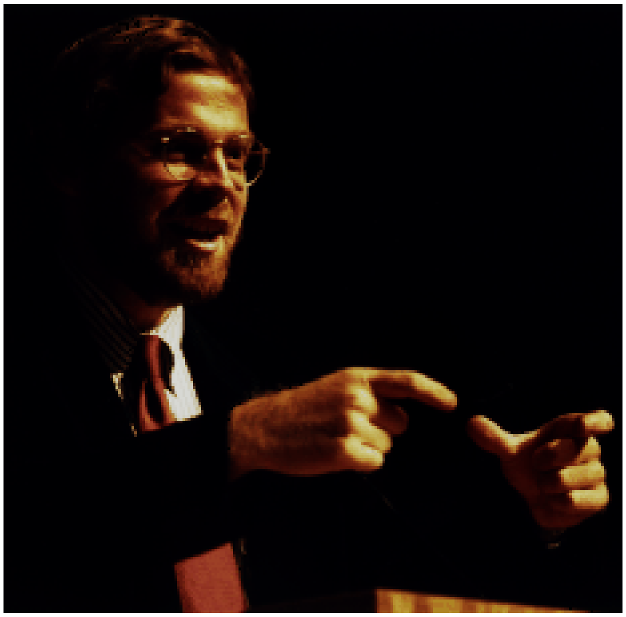

Figure 1: Hawley talking about mosaic tiles

Gestures and facial displays such as those demonstrated by Mike Hawley could be profitably employed in human-computer dialogue systems as well. Let's deconstruct exactly what people do with their hands and faces during dialogue, and how the function of the three modalities are related.

Nonverbal Behaviors in Human Dialogue

Kinds of Gesture

Emblems

We’re going to start by looking at a gesture-type that has not proved particularly useful to automatic dialogue systems. And yet, when we reflect on what kinds of gestures we have seen in our environment, we often come up with exactly that type of gesture, known as *emblematic* . These gestures are culturally specified in the sense that one single gesture may differ in interpretation from culture to culture (Efron, 1941; Ekman & Friesen, 1969). For example, the American “V-for-victory” gesture can be made either with the palm **or** the back of the hand towards the listener. In Britain, however, a ‘V’ gesture made with the back of the hand towards the listener is inappropriate in polite society. Examples of emblems in American culture are the thumb-and-index-finger ring gesture that signals 'okay' or the 'thumbs up' gesture. Many more of these "emblems" appear to exist in French and Italian culture than in America (Kendon, 1993), but in few cultures do these gestures appear to constitute more than 10% of the gestures produced by speakers. Despite the paucity of emblematic gestures in everyday communication, it is gestures such as these that have interested computer scientists. That is, computer vision systems known as “gestural interfaces” attempt to invent or co-opt emblematic gesture to replace language in human-computer interaction. However, in terms of *types*, few enough different emblematic gestures exist to make the idea of co-opting emblems as a gestural language untenable. And in terms of *tokens*, we simply don’t seem to make that many emblematic gestures on a daily basis. In dialogue systems, then, where speech is already a part of the interaction, it makes more sense to concentrate on integrating those gestures that accompany speech in human-human conversation.

Propositional Gestures

Another conscious gesture that has been the subject of some study in the interface community is the so-called 'propositional gesture' (Hinrichs & Polanyi, 1986). An example is the use of the hands to measure the size of a symbolic space while the speaker says "it was this big". Another example is pointing at a chair and then pointing at another spot and saying “move that over there”. These gestures are not unwitting and in that sense not spontaneous, and their interaction with speech is more like the interaction of one grammatical constituent with another than the interaction of one communicative channel with another; in fact, the demonstrative "this" may be seen as a place holder for the syntactic role of the accompanying gesture. These gestures can be particularly important in certain types of task-oriented talk, as discussed in the well-known paper “Put-That-There: Voice and Gesture at the Graphics Interface” (Bolt, 1980). Gestures such as these are found notably in communicative situations where the physical world in which the conversation is taking place is also the topic of conversation. These gestures do not, however, make up the majority of gestures found in spontaneous conversation, and I believe that in part they have received the attention that they have because they are, once again, *conscious witting* gestures available to our self-scrutiny.

Spontaneous Gestures

Let us turn now to the vast majority of gestures; those that although unconscious and unwitting are the gestural vehicles for our communicative intent with other humans, and potentially with our computer partners as well. These gestures, for the most part, are not available to conscious access, either to the person who produced them, or to the person who watched them being produced. The fact that we lose access to the form of a whole class of gestures may seem odd, but consider the analogous situation with speech. For the most part, in most situations, we lose access to the *surface structure* of utterances immediately after hearing or producing them (Johnson et al., 1973). That is, if listeners are asked whether they heard the word “couch” or the word “sofa” to refer to the same piece of furniture, unless one of these words sounds odd to them, they probably will not be able to remember which they heard. Likewise, slight variations in pronunciation of the speech we are listening to are difficult to remember, even right after hearing them (Levelt, 1989). That is because (so it is hypothesized), we listen to speech in order to extract meaning, and we throw away the words once the meaning has been extracted. In the same way, we appear to lose access to the form of gestures (Krauss, Morrel-Samuels & Colasante, 1991), even though we attend to the information that they convey (Cassell et al, 1998).

The spontaneous unplanned, more common **co-verbal** gestures are of four types:

- *Iconic* gestures depict by the form of the gesture some feature of the action or event being described. An example is a gesture outlining the two sides of a triangle while the speaker said "the biphasic-triphasic distinction between gestures is the first cut in a hierarchy".

Iconic gestures may specify the viewpoint from which an action is narrated. That is, gesture can demonstrate who narrators imagine themselves to be, and where they imagine themselves to stand at various points in the narration, when this is rarely conveyed in speech, and listeners can infer this viewpoint from the gestures they see. For example, a participant at a computer vision conference was describing to his neighbor a technique that his lab was employing. He said "and we use a wide field cam to [do the body]'", while holding both hands open and bent at the wrists with his fingers pointed towards his own body, and the hands sweeping up and down. His gesture shows us the wide field cam "doing the body", and takes the perspective of somebody whose body is "being done". Alternatively, he might have put both hands up to his eyes, pantomiming holding a camera, and playing the part of the viewer rather than the viewed.

- *Metaphoric gestures* are also representational, but the concept they represent has no physical form; instead the form of the gesture comes from a common metaphor. An example is the gesture that a conference speaker made when he said "we're continuing to expound on this" and made a rolling gesture with his hand, indicating ongoing process.

Some common metaphoric gestures are the 'process metaphoric' just illustrated, and the 'conduit metaphoric' which objectifies the information being conveyed, representing it as a concrete object that can be held between the hands and given to the listener. Conduit metaphorics commonly accompany new segments in communicative acts; an example is the box gesture that accompanies "In this [next part] of the talk I'm going to discuss new work on this topic". Metaphoric gestures of this sort contextualize communication; for example, placing it in the larger context of social interaction. In this example, the speaker has prepared to give the next segment of discourse to the conference attendees. Another typical metaphoric gesture in academic contexts is the metaphoric pointing gesture that commonly associates features with people. For example, during a talk on spontaneous gesture in dialogue systems, I might point to Phil Cohen in the audience while saying “I won’t be talking today about the pen gesture”. In this instance I am associating Phil Cohen with his work on pen gestures.

- *Deictics* spatialize, or locate in the physical space in front of the narrator, aspects of the discourse; these can be discourse entities that have a physical existence, such as the overhead projector that I point to when I say “this doesn’t work”, or non-physical discourse entities. An example of the latter comes from an explanation of the accumulation of information during the course of a conversation. The speaker said “we have an [attentional space suspended] between us and we refer [back to it]”. During “attentional space” he defined a big globe with his hands, and during “back to it” he pointed to where he had performed the previous gesture.

Deictic gestures populate the space in between the speaker and listener with the discourse entities as they are introduced and continue to be referred to. Deictics do not have to be pointing index fingers. One can also use the whole hand to represent entities or ideas or events in space. In casual conversation, a speaker said “when I was in a [university] it was different, but now I’m in [industry]” while opening his palm left and then flipping it over towards the right. Deictics may function as an interactional cue, indexing which person in a room the speaker is addressing, or indexing some kind of agreement between the speaker and a listener. An example is the gesture commonly seen in classrooms accompanying “yes, [student X], you are exactly right” as the teacher points to a particular student.

- Beat gestures are small baton like movements that do not change in form with the content of the accompanying speech. They serve a pragmatic function, occurring with comments on one's own linguistic contribution, speech repairs and reported speech.

Beat gestures may signal that information conveyed in accompanying speech does not advance the "plot" of the discourse, but rather is an evaluative or orienting comment. For example, the narrator of a home repair show described the content of the next part of the TV episode by saying "I'm going to tell you how to use a caulking gun to [prevent leakage] through [storm windows] and [wooden window ledges]. . ." and accompanied this speech with several beat gestures to indicate that the role of this part of the discourse was to indicate the relevance of what came next, as opposed to imparting new information in and of itself. Beat gestures may also serve to maintain conversation as dyadic: to check on the attention of the listener, and to ensure that the listener is following (Bavelas et al., 1992).

The importance of these four types of gestures is that (a) they convey content that is not conveyed by speech, (b) their placement tells us something about the state of the conversation, (c) their placement tells us something about the structure of the discourse. In addition, the fact that they convey information that is not conveyed by speech gives the impression of cognitive activity over and above that required for the production of speech. That is, (d) they give the impression of a *mind*, and therefore, when produced by embodied dialogue systems, enhance the believability of the interactive system. To be able to exploit any of these four properties in the construction of embodied spoken dialogue systems requires an understanding of the *integration* of gesture with speech. This is what we turn to next.

Integration of Gesture with Spoken Language

Gestures are integrated into spoken dialogue at the level of the phonology, the semantics, and discourse structure.

Temporal Integration of Gesture and Speech

First, a short introduction to the physics of gesture: iconic and metaphoric gestures are composed of three phases. And these *preparation*, *stroke*, and *retraction* phases may be differentiated by short holding phases surrounding the stroke. Deictic gestures and beat gestures, on the other hand, are characterized by two phases of movement: a movement into the gesture space, and a movement out of it. In fact, this distinction between biphasic and triphasic gestures appears to correspond to the addition of semantic features—or iconic meaning—to the representational gestures. That is, the number of phases corresponds to type of meaning: representational vs. non-representational. And it is in the second phase—the stroke—that we look for the meaning features that allow us to interpret the gesture (Wilson, Bobick & Cassell, 1996). At the level of the word, in both types of gestures, individual gestures and words are synchronized in time so that the 'stroke' (most energetic part of the gesture) occurs either with or just before the intonationally most prominent syllable of the accompanying speech segment (Kendon, 1980; McNeill, 1992).

This phonological co-occurrence leads to co-articulation of gestural units. Gestures are performed rapidly, or their production is stretched out over time, so as to synchronize with preceding and following gestures, and the speech these gestures accompany. An example of gestural co-articulation is the relationship between the two gestures in the phrase “do you have an ACCOUNT at this BANK?”: during the word “account”, the two hands sketch a kind of box in front of the speaker; however, rather than carrying this gesture all the way to completion (either both hands coming to rest at the end of this gesture, or maintaining the location of the hands in space), one hand remains in the 'account' location while the other cuts short the 'account' gesture to point at the ground while saying 'bank'. Thus, the occurrence of the word “bank”, with its accompanying gesture, affected the occurrence of the gesture that accompanied “account”.

At the level of the turn, the hands being in motion is one of the most robust cues to turn-taking (Duncan, 1974). Speakers bring their hands into gesture space as they think about taking the turn, and at the end of a turn the hands of the speaker come to rest, before the next speaker begins to talk. Even clinical stuttering, despite massive disruptions of the flow of speech, does not interrupt speech-gesture synchrony. Gestures during stuttering bouts freeze into holds until the bout is over, and then speech and gesture resume in synchrony (Scoble 1993). In each of these cases, the linkage of gesture and language strongly resists interruption.

Semantic Integration

Speech and the non-verbal behaviors that accompany it are sometimes redundant, and sometimes they present complementary but non-overlapping information. This complementarity can be seen at several levels.

In the previous section I said that gesture is cotemporaneous with the linguistic segment it most closely resembles in meaning. But what meanings does gesture convey, and what is the relationship between the meaning of gesture and of speech? Gesture can convey redundant or complementary meanings to those in speech—in normal adults gesture is almost never contradictory to what is conveyed in speech (politicians may be a notable exception, if one considers them normal adults). At the semantic level this means that the semantic features that make up a concept may be distributed across speech and gesture. As an example, take the semantic features of manner of motion verbs: these verbs, such as “walk”, “run”, “drive”, can be seen as being made up of the meaning “go” **plus** the meanings of how one got there (walking, running, driving). The verbs “walking” and “running” can be distinguished by way of the speed with which one got there. And the verb “arrive” can be distinguished from “go” by whether one achieved the goal of getting there, and so on. These meanings are semantic features that are added together in the representation of a word. Thus, I may say “he drove to the conference” or “he went to the conference” + drive gesture. McNeill has shown that speakers of different languages make different choices about which features to put in speech and which in gesture (McNeill, forthcoming). Speakers of English often convey path in gesture and manner in speech, while speakers of Spanish put manner in gesture and path in speech. McNeill claims that this derives from the typology of Spanish vs. English.

In my lab, we have shown that even in English a whole range of features can be conveyed in gesture, such as path, speed, telicity ("goal-achievedness"), manner, aspect. One person, for example, said "Road Runner [comes down]" while she made a gesture with her hands of turning a steering wheel. Only in the gesture is the manner of coming down portrayed. She might just have well as said "Road Runner comes down" and made a walking gesture with her hands. Another subject says "Road Runner just [goes]" and with one index finger extended makes a fast gesture forward and up, indicating that the Road Runner zipped by. Here both the path of the movement (forward and up) and the speed (very fast) are portrayed by the gesture, but the manner is left unspecified (we don't know whether the Road Runner walked, ran, or drove).

Even among the blind, semantic features are distributed across speech and gesture—strong evidence that gesture is a product of the same generative process that produces speech. Children who have been blind from birth and have never experienced the communicative value of gestures do produce gestures along with their speech (Iverson & Goldin-Meadow, 1996). The blind perform gestures during problem solving tasks, such as the Piagetian conservation task. Trying to explain why the amount of water poured from a tall thin container into a short wide container is the same (or is different, as a non-conserver would think), blind children, like sighted ones, perform gestures as they speak. For example, a blind child might say "this one was tall" and make a palm-down flat-hand gesture well above the table surface, and say "and this one is short" and make a two handed gesture indicating a short wide dish close to the table surface. Only in the gesture is the wide nature of the shorter dish indicated.

Discourse Integration

For many gestures, occurrence is determined by the discourse structure of the talk. In particular, information structure appears to play a key role in the distribution of gesture. The information structure of an utterance defines its relation to other utterances in a discourse and to propositions in the relevant knowledge pool. Although a sentence like "George withdrew fifty dollars" has a clear semantic interpretation which we might symbolically represent as *withdrew'(george', fifty-dollars')* , such a simplistic representation does not indicate how the proposition relates to other propositions in the discourse. For example, the sentence might be an equally appropriate response to the questions “Who withdrew fifty dollars”, “What did George withdraw”, “What did George do”, or even “What happened”? Determining which items in the response are most important or salient clearly depends on which question is asked. These types of salience distinctions are encoded in the information structure representation of an utterance.

Following Halliday and others (Halliday 1967; Hajicova & Sgall, 1987), we use the terms *theme* and *rheme* to denote two distinct information structural attributes of an utterance. The theme/rheme distinction is similar to the distinctions *topic/comment* and *given/new*. The theme roughly corresponds to what the utterance is about, as derived from the discourse model. The rheme corresponds to what is new or interesting about the theme of the utterance. Depending on the discourse context, a given utterance may be divided on semantic and pragmatic grounds into thematic and rhematic constituents in a variety of ways. That is, depending what question was asked, the contribution of the current answer will be different[2].

In English intonation serves an important role in marking information as rhematic, and in marking it as contrastive[3]. That is, pitch accents mark which information is new to the discourse. Thus, the following two examples demonstrate the association of pitch accents with information structure (primary pitch accents are shown in bold face type):

[Q:] Who withdrew fifty dollars?
[A:] **(George)** $_{\text{RHEME}}$ (withdrew fifty dollars) $_{\text{THEME}}$

[Q:] What did George withdraw?
[A:] (George withdrew)$_{\text{THEME}}$ **(fifty dollars)**$_{\text{RHEME}}$

[2] This description is, of course, an oversimplification of a topic that is still the subject of vigorous debate. We do not pretend to a complete theory which would, in any case, be beyond the scope of the current chapter.

[3] See Hirschberg, this volume, for further details about the role of intonation in spoken dialogue.

In speaking these sentences aloud one notices that even though the answers to the two questions are identical in terms of the words they contain, they are uttered quite differently: in the first the word "George" is stressed, and in the second it is the phrase "fifty dollars" which is stressed. This is because in the two sentences different elements are marked as rhematic, or difficult for the listener to predict .

Gesture also serves an important role in marking information structure. When gestures are found in an utterance, the vast majority of them co-occur with the rhematic elements of that utterance (Cassell & Prevost, in preparation). In this sense intonation and gesture serve similar roles in the discourse. Intonational contours also time the occurrence of gesture (Cassell & Prevost, op. cit.). Thus, the distribution of gestural units in the stream of speech is similar to the distribution of intonational units, in the following ways.

- First, gestural domains are isomorphic with intonational domains. The speaker's hands rise into space with the beginning of the intonational rise at the beginning of an utterance, and the hands fall at the end of the utterance along with the final intonational marking.
- Secondly, the most effortful part of the gesture (the "stroke") co-occurs with the pitch accent, or most effortful part of enunciation.
- Third, gestures co-occur with the rhematic part of speech, just as we find particular intonational tunes co-occurring with the rhematic part of speech. We hypothesize that this is so because the rheme is that part of speech that contributes most to the ongoing discourse, and that is least known to the listener beforehand. It makes sense that gestures, which may convey additional content to speech and may flag that part of the discourse as meriting further attention, would be found where the most explanation is needed in the discourse. This does not mean that one never finds gestures with the theme, however. Some themes are *contrastive*, marking the contrast between one theme and another. An example is "In the cartoon you see a manhole

cover. And then the rock falls down **on that manhole cover**". When thematic material is contrastive, then gesture may occur in that context.

In sum, then, gestures of four types co-occur with speech in particular rule-governed ways. These associations serve to mark the status of turn-taking, to mark particular items as rhematic (particularly important to the interpretation of the discourse), and to convey meanings complementary to those conveyed by speech. Are these results true only for North America?

Cultural Differences

It is natural to wonder about the cultural specificity of gesture use. We often have the impression that Italians gesture more and differently than do British speakers. It is true that, as far as the question of quantity is concerned, that speakers from some language communities demonstrate a greater number of gestures per utterance than others, a phenomenon which appears to be linked to the fact that some cultures may embrace the use of gesture more than others—many segments of British society believe that gesturing is inappropriate, and therefore children are encouraged to not use their hands when they speak. But the effect of these beliefs and constraints about gesture is not as strong as one might think. In my experience videotaping people carrying on conversations and telling stories, many speakers claim that they never use their hands. These speakers are then surprised to watch themselves on video, where they can be seen using their hands as much as the next person. In fact, every speaker of every language that I have seen videotaped (French, Spanish, Italian, Tagalog, Filipino, Soviet Georgian, Chinese, Japanese, to name a few) has gestured . . . all except for one American man who made one single gesture during his entire 20 minutes narration, a gesture which he himself aborted by grabbing the gesturing hand with the other hand and forcefully bringing it down to his lap.

As far as the nature of gesture is concerned, as mentioned above, emblems do vary widely from language community to language community. Americans make a 'V for Victory' with their palm oriented either out towards the listener or towards themselves. For British speakers, the ‘V for Victory with the palm oriented towards the self is exceedingly rude. Italian speakers demonstrate a wide variety of emblematic gestures that can carry meaning in the absence of speech, while both American and English speakers have access to a limited number of such gestures. But remember that emblematic gestures still make up less than 20% of the gestures found in everyday conversation. The four gesture types of spontaneous gestures described, however, appear universal. Interestingly, and perhaps not surprisingly, the *form* of metaphoric gestures appears to differ from language community to language community. Conduit metaphoric gestures are not found in narrations in all languages: neither Chinese nor Swahili narrators use them, for example (McNeill, 1992). These narratives do contain abundant metaphoric gestures of other kinds, but do not depict abstract ideas as bounded containers. The metaphoric use of space, however, appears in all narratives collected, regardless of the language spoken. Thus, apart from emblematic gestures, the use of gesture appears to be more universal than particular.

Kinds of Facial Displays

Let us turn now to the use of the face during conversation. Like hand gestures, facial displays can be classified according to their placement with respect to the linguistic utterance and their significance in transmitting information (Scherer, 1980). When we talk about facial displays we are really most interested in precisely-timed changes in eyebrow position, expressions of the mouth, movement of the head and eyes, and gestures of the hands. For example, raised eyebrows + a smiling mouth, is taken to be a happy expression (Ekman & Friesen, 1984), while moving one's head up and down is taken to be a nod. Some facial displays are linked to personality, and remain constant across a lifetime (a "wide-eyed look"), some are linked to emotional state, and may last as long as the emotion is felt (downcast eyes during depression), and some are synchronized with the units of conversation and last only a very short time (eyebrow raises along with pitch-accented words).

As well as characterizing facial displays by the muscles or part of the body in play, we can also characterize them by their function in a conversation. Some facial displays have a phonological function—lip shapes that change with the phonemes uttered. Although it has been shown that such lip shapes can significantly improve the facility with which people understand “talking heads” (Bregler et al, 1993), this function of facial displays will not be further discussed in the current chapter. Some facial displays fulfill a semantic function, for example nodding rather than saying “yes”. Some facial displays convey transitory emotional states (emotional feedback to the conversation). Examples include smiling when asked if one would like ice-cream or looking puzzled when queried about one's non-existent sister. Some facial displays, on the other hand, have an envelope[4], or conversational-process oriented function. Examples are quick nods of the head while one is listening to somebody speak, a glance at the other person when one is finished speaking, or a beat gesture when one is taking the turn. Still other functions for facial displays are to cement social relationships (polite smiles), and to correspond to grammatical functions (eyebrow raises on pitch-accented words).

Note that the same movements by the body can have two (or more) different functions. Smiles can serve the function of emotional feedback, indicating that one is happy, or they can serve a purely social function even if one is not at all happy. Nods of the head can replace saying "yes" (a content function), or simply indicate that one is following, even if one does not agree with what is being said (an envelope function).[5]

[4] We call these behaviors "envelope" to convey the fact that they concern the outer envelope of communication, rather than its contents. A similar distinction between content and envelope is made by Takeuchi & Nagao (1993) when they refer to the difference between "object-level communication" (relevant to the communication goal) and "meta-level processing" (relevant to communication regulation).

[5] Content nods tend to be fewer and produced more emphatically and more slowly than envelope nods (Duncan, 1974).

Cultural Differences

Note that, like emblem gestures, facial displays with a semantic function can vary from culture to culture. To indicate agreement, for example, one nods in the United States but shakes one's head in Greece or Albania. However, like beat gestures, facial displays with a dialogic function are similar across cultures. Thus, although generally one looks less often at one's interlocutor in Japanese conversation, conversational turns are still terminated by a brief glance at the listener. And, even though semantic agreement is indicated by a shake of the head in Greece, feedback is still accomplished by a nod. As with gesture, then, there are more similarities in the use of the face than there are differences, at least with respect to the regulatory conversational function of these behaviors.

In the remainder of this chapter, we concentrate on the nonverbal behaviors whose description has universal validity.

Integration of Verbal Displays with Spoken Language

As with gesture, facial displays are tightly coupled to the speech that they occur with.

Temporal synchronization

When a word is in the process of being articulated, eye blinks, hand movement, head turning, brow raising occur, and they finish at the end of the word. Synchrony occurs at all levels of speech: phonemic segment, word, phrase or long utterance. Different facial motions are isomorphic to these groups (Condon & Osgton, 1971; Kendon, 1974). Some of them are more adapted to the phoneme level, like an eye blink, while others occur at the word level, like a frown. In the example "Do you have a checkbook with you?", a raising eyebrow starts and ends on the accented syllable "check"; while a blink starts and ends on the pause marking the end of the utterance. Facial display of emphasis can match the emphasized segment, showing synchronization at this level (a sequence of head nods can punctuate the emphasis, as when one nods while saying the word ‘really’ in the phrase “I REALLY want this system to work”). Moreover, some movements reflect encoding-decoding difficulties and therefore coincide with hesitations and pauses inside clauses (Dittman, 1974). Many hesitation pauses are produced at the beginning of speech, and correlate with avoidance of gaze, presumably to help the speaker concentrate on what s/he is going to say.

Facial Display Occurrence

As described above, facial displays can be classified according to their placement with respect to the linguistic utterance and their significance in transmitting information. Some facial displays have nothing to do with the linguistic utterance, but serve a biological need (wetting the lips or blinking), some are synchronized with phonemes, such as changing the shape of one’s lips to utter a particular sound. The remaining facial displays, that have a dialogic function, are primarily movements of the eyes (gaze), eyebrow raises, and nods. In the following section we discuss the co-occurrence of these behaviors with the verbal utterance.

Facial behavior can be classified into four primary categories depending on its role in the conversation (Argyle & Cook, 1976; Collier, 1985). The following describes where behaviors in each of these four categories occur, and how they function.

- <u>Planning</u>: planning eye movements correspond to the first phase of a turn when speakers organize their thoughts. The speaker has a tendency to look away in order to prevent an overload of visual and linguistic information . On the other hand, during the execution phase, when speakers know what they are going to say, they look more often at listeners. For a short turn (of a duration of less than 1.5 seconds), this planning look-away does not occur, and the speaker and listener maintain mutual gaze.

- Comment: accented or emphasized linguistic items are punctuated by head nods; the speaker may also look toward the listener at these moments. Eyebrow raises are also synchronized with pitch accents.
- Control: Some eye movements regulate the use of the communication channel and function as synchronization signals. That is, one may request a response from a listener by looking at the listener, and suppress the listener's response by looking away; these behaviors occur primarily at the ends of utterances and at grammatical boundaries. When the speaker wants to give up the floor, she gazes at the listener at the end of the utterance. When the listener wants the floor, she looks at and slightly up at the speaker.
- Feedback: Facial behaviors may be used to elicit feedback and to give it. Speakers look toward listeners during grammatical pauses and when asking questions, and these glances signal requests for verbal or non-verbal feedback, without turning the floor over to the listener. Listeners respond by establishing gaze with the speaker and/or nodding. The feedback-elicitation head movements are referred to as *within turn* signals . If the speaker does not emit such a signal by gazing at the listener, the listener can still emit a *back-channel* or feedback signal, which in turn may be followed by a *continuation* signal by the speaker. But the listener's behavior is dependent on the behavior of the speaker; one is much less likely to find feedback in the absence of a feedback elicitation signal (Duncan, 1974).

In the description just given, facial behavior is described as a function of the turn-taking structure of a conversation rather than as a function of information structure. This is the way that facial behavior has been described in the majority of the literature; in fact, gaze behavior has come to represent *the* cue to turn-organization and has been described as if it were entirely predicted by turn-organization. However, turn-taking only partially accounts for the gaze behavior in discourse. Our research shows that a better explanation for gaze behavior integrates turn-taking with the information structure of the propositional content of an utterance (Cassell, Torres & Prevost, in press). Specifically, the beginning of themes are frequently accompanied by a look-away from the hearer, and the beginning of rhemes are frequently accompanied by a look-toward the hearer. When these categories are co-temporaneous with turn-construction, then they are strongly—in fact, absolutely—predictive of gaze behavior. That is, when the end of the rheme corresponds to the end of the turn, then speakers always look towards their listeners in our data.

Why might there be such a link between gaze and information structure? The literature on gaze behavior and turn-taking suggests that speakers look toward hearers at the ends of turns to signal that the floor is "available"—that hearers may take the turn. Our findings suggest that speakers look toward hearers at the beginning of the rheme—that is, when new information or the key point of the contribution is being conveyed. Gaze here may focus the attention of speaker and hearer on this key part of the utterance. And, of course, signaling the new contribution of the utterance and signaling that one is finished speaking are not entirely independent. Speakers may be more likely to give up the turn once they have conveyed the rhematic material of their contribution to the dialogue. In this case, gaze behavior is signaling a particular kind of relationship between information structure and turn-taking. It is striking, both in the role of facial displays in turn-taking and in their association with information structure, the extent to which these 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 fill the function that Brennan & Hulteen (1995) suggest is needed for more robust speech interfaces.

Two Demonstrations

Before turning to the ways in which we have deployed these rules for the integration of verbal and non-verbal behavior in embodied dialogue systems, I'm going to present two fragments of dialogue showing the rules in action. The first example is taken from life (another one of my colleagues) while the second (from Cassell et al., 1994) is, in fact, a conversation that took place between two computer-generated animated agents, and in which all verbal and non-verbal behaviors were generated automatically, by rule.

Figure 2 shows Seymour Papert 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.

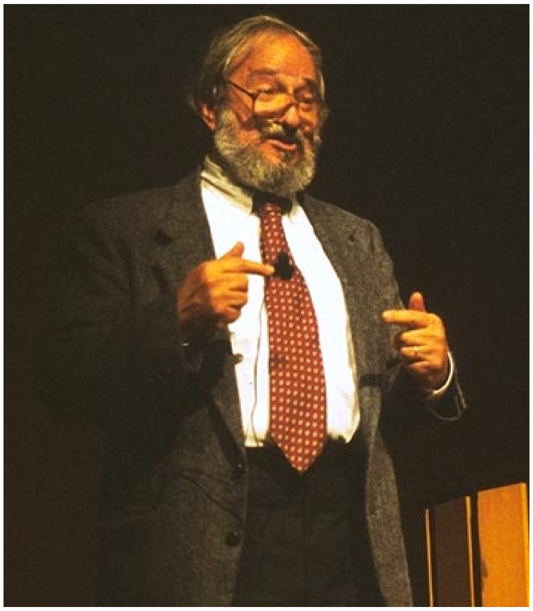

Figure 2: ". . . will interact"

Now, because this is a speech, and not a conversation, some of the integration with verbal and nonverbal behavior is different, 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 still uses all four kinds of gestures (beat gestures, in fact, although not illustrated here, are particularly frequent in speeches). 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, perseverates through his speech disfluency, allowing the second gesture, an iconic, 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 began to 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.

Before we turn to a description of our embodied conversational agents, let's start with an example of their behavior. The system that I will describe here is not, strictly speaking, a dialogue system, because it talks to another embodied agent, and not to a human. But the example is closer to what we would like to achieve (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.

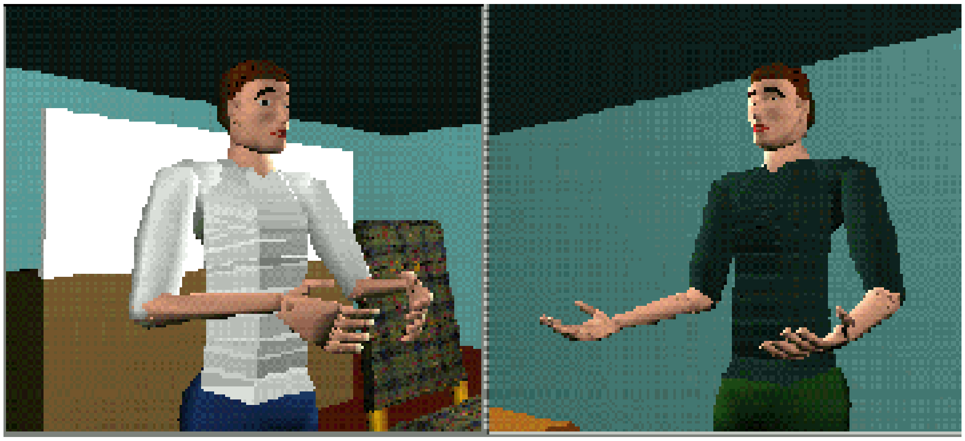

Figure 3: (a) "do you have a blank check?"; (b) “can you help me?”

Figure 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”.

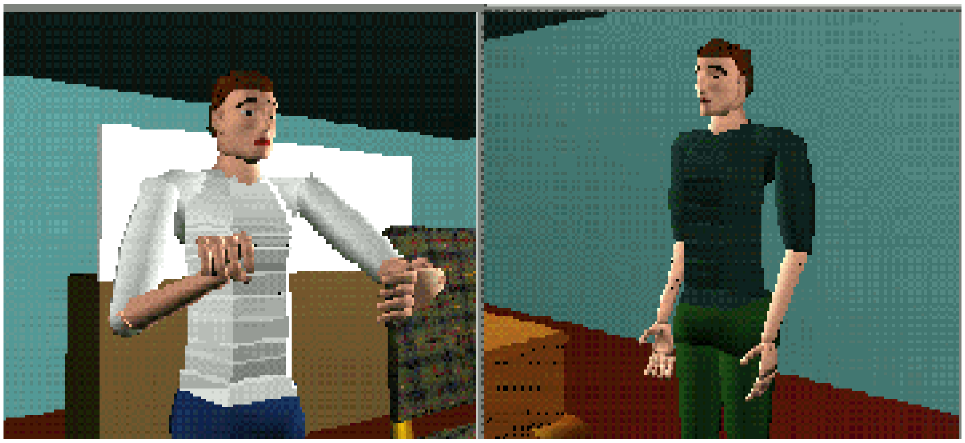

Figure 4: (a) "you can write the check"; (b) “I have eighty dollars”

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 (Badler et al., 1993) as the body. Our focus in this 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 (1977), 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.

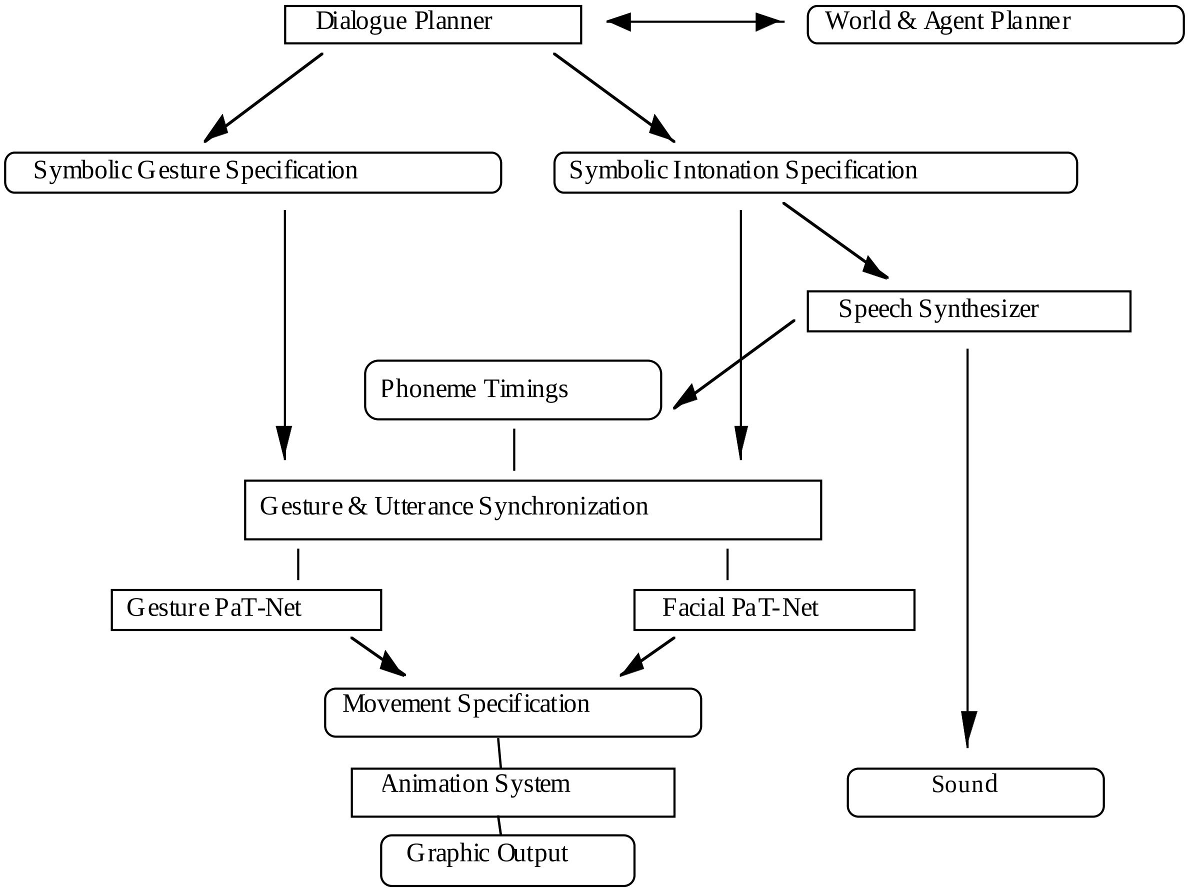

Figure 5: Architecture of Animated Conversation

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 his own and ultimately returns his next contribution to their discourse. Depending on this message received in response, the agent may have to modify his 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 which 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 (1991) and Prevost & Steedman (1994). 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 (Liberman & Pierrehumbert, 1984). 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 (concepts such as 'checkbook', 'write', etc.).
- Concepts with common metaphoric vehicles received metaphorics (concepts such as 'withdraw [money]', 'bank account', 'needing help');
- Concepts referring to places in space received deictics ('here', 'there').
- 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].

The timing of gestures was also implemented according to the rules I described above. Information about the duration of intonational phrases was acquired during speech generation and then used to time gestures. If there was a non-beat gesture in an utterance, its preparation was set to begin at or

[6] The lexicon look-up approach was a temporary solution to the problem of specifying the semantic content of gestures, which will be addressed further below.

before the beginning of the intonational phrase, and to finish at or before the first beat gesture in the intonational phrase or the nuclear stress of the phrase, whichever came first. The stroke phase was set to coincide with the nuclear stress of the phrase. Finally, the relaxation was set to begin no sooner than 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, were simply timed to coincide with the stressed syllable of the word that realized the associated concept.

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

Similarly, facial displays were generated as a function of the dialogue 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

[7] For more details on the facial animation, see Pelachaud et al., 1994

text and graphics (Wahlster et al., 1991), and multimodal integration in input (Johnston et al., 1997). 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 built 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 understand human communicative behavior, and generate integrated spontaneous verbal and nonverbal behavior of their own (see Thórisson, 1994, forthcoming 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. Ymir offers thus 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. 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 (Duncan, 1974). 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 which 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.

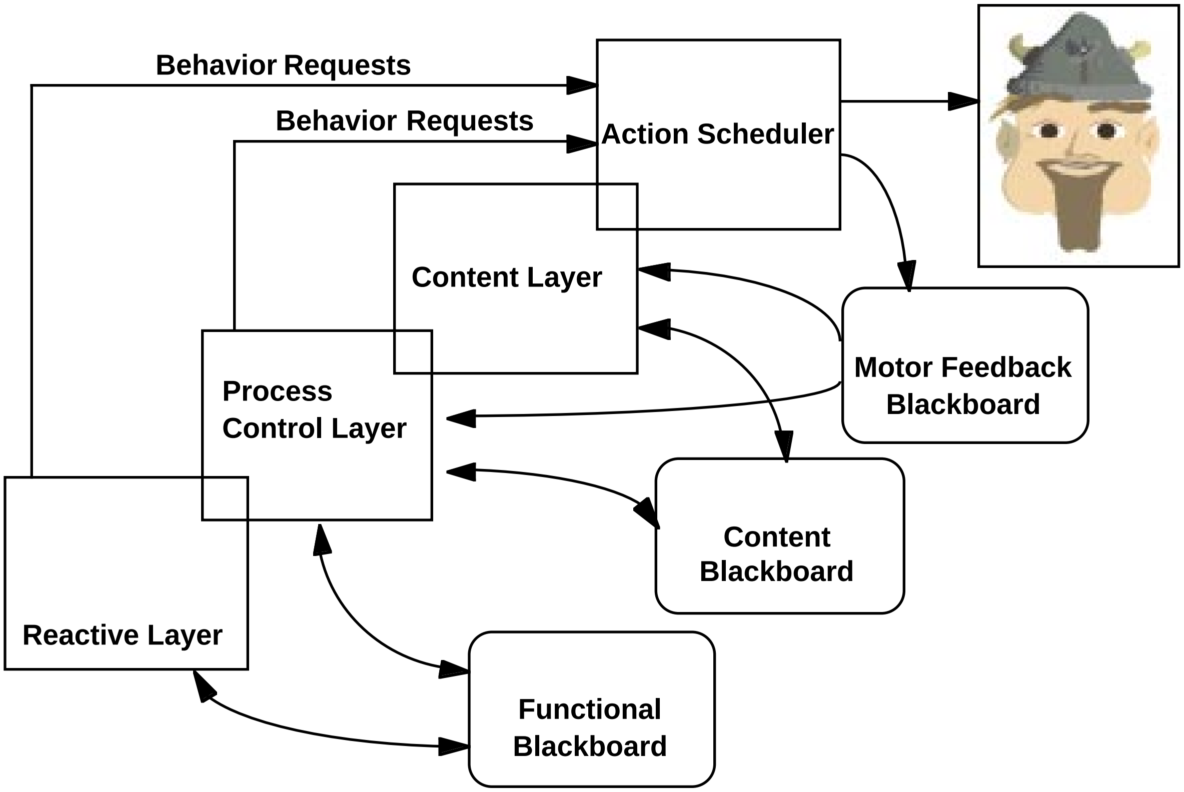

Figure 6: Gandalf, blackboard architecture

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

[8] For a more complete treatment of the role of body orientation and other whole body behaviors in conversation, see our BodyChat system, a 3D-graphical world in which the conversational body behaviors of avatars are automatically generated as a function of stage of the conversation, social distance, etc. (Vilhjálmsson & Cassell, 1998).

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 understands gestures pointing towards the screen as references to planets, and also understands 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, he is able to converse with people and to understand 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 he 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 him 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, still under construction, 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 our plans for the Rea system, which is partially constructed as of this date.

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 will require a trade off between linguistic processing and generation, and reactivity. Our goal, then, is to design Rea in such a way that her reactions are aptly timed, and may even provide more time for the reflective layer to come up with the correct content, either in understanding 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

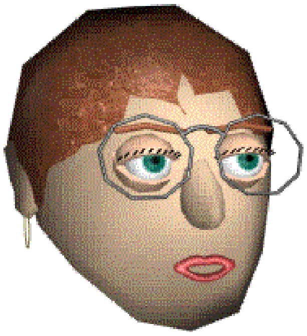

Figure 7: Rea, the real estate agent

other words, Rea will be constructed from a suite of communication skill processes which are operating with different response times.

The domain which we have chosen initially is real estate: Rea will be capable of interacting 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 . Rea will be able to engage in social chit-chat, and she will be able to take users through 3D walk-throughs of different houses on a large screen. She will be able both to answer questions about particular houses, and to initiate conversation .

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 in non-trivial, mixed-initiative, multimodal dialogue, 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 understanding. In a multimodal conversation system, the understanding 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.

Evaluation: Do Bodies Offer Anything to Dialogue Systems?

Of the two systems that are fully implemented, Gandalf has lent itself more readily to evaluation. That is, because of Gandalf/Ymir's modular architecture, it is quite easy to build agents with different kinds of conversational skills. This flexibility allow 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 (Ekman, 1979). 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. Elliott, 1997, Koda & Maes, 1996, Nagao & Takeuchi, 1994, Takeuchi & Nagao, 1993). 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 & Thorisson, forthcoming).

Figure 8: Preparing to interact with Gandalf

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

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, he gave a confused expression when he didn’t understand an utterance, and he 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 his 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 he 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 his head and eyes towards the user when listening, and towards task when executing commands in the domain. He could avert his gaze and lift his eyebrows when taking turn. He gazed at the person when giving turn. Finally, he 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 he brings it up on the screen, then turns to user and speaks*]
User: "Tell me more." [*Gandalf takes about 2 seconds to parse the speech, but he knows within 250 ms when the user gives the turn, so he looks to the side to show that he's taking the turn, and his eyebrows go up and down as he hesitates 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 (Cassell et al., 1998) 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 (Cassell & Prevost, in preparation).

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 & O'Conaill (1997) 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. In this section I discuss the directions that my students and I are taking in this arena. Let’s look first at the issue of generating gestural form from scratch. 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. In this way we can implement in gesture generation the insights I described above about the role of gesture in describing motion. 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.

In order to determine *when* content is distributed across speech and gesture, and when it is conveyed *redundantly* in both speech and gesture, Prevost and I have begun to look at data from human-human conversation (Cassell & Prevost, in preparation). We have found that, as we implemented in Animated Conversation, gestures are overwhelmingly found in association with the rheme of an utterance. Within the rheme, however, the question of redundancy is mediated by whether iconic gestures represent information from the point of view of an observer, or the point of view of the speaker (see section on Spontaneous Gestures, above). When gestures portray the point of view of an observer, then gestures and speech are almost always redundant. When gestures portray the point of view of a character, or the speaker, then the gesture often conveys information that is not conveyed in accompanying speech. Results such as these are allowing us to refine the determination of the form and placement of gestures in association with spoken language.

Finally, it should be noted that, unlike what has been posited for humans, the generation of gesture and speech in the systems built to date has been quite linear, without any chance of feedback between the modalities. To address this issue, Scott Prevost and I have been thinking about a way for gesture and intonation to affect one another. Prevost (1996) argues that the determination of focus (and hence pitch accent placement) within thematic and rhematic (old information or new information) constituents should be handled by the sentence planner. Based on this observation and the mapping of triphasic gestures onto intonational tunes described in Cassell et al. (1994), we can also assert that the alignment of the three gesture phases with the intonation contour occurs at this level as well. This aspect of our architecture has a strong effect on the interaction between speech and gesture in generation: the choice of gestures and choice of speech form interact such that gesture will actually affect where stress is placed in the utterance. For example, if a sentence such as “Road Runner zipped over Coyote” is planned then, depending on the gesture chosen, as well as the underlying representation, primary stress will be differently assigned. If the gesture chosen represents driving, then primary stress will fall on “zipped” (as the reader can see by reading the sentence out loud, it is difficult to imagine performing the gesture along with “over”, or stressing the word “over” if the gesture co-occurs with “zipped”). If, on the other hand, the gesture chosen simply represents motion from point A to point B, then primary stress might fall on “zipped” or on “over” depending which of these terms is focused (or contrastive) in the context of the text. This is an exciting direction to pursue because it means that the generation of one modality can have an effect on the generation of other modalities.

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. (1997)’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 (1997) 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, around 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 & Takeuchi (1993), 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 & Badler (1997) 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 (1997). Lester et al. (forthcoming) 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 (1996) 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 & Hulteen have pointed out (1995), conversation is fundamentally collaborative, and dialogue systems can be improved by focusing 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

André, E., Rist, T. and Mueller, J. (forthcoming). Employing AI methods to control the behavior of animated interface agents. *Applied Artificial Intelligence.*

Badler, N., Phillips, C. and Webber, B. (1993). *Simulating Humans: Computer Graphics Animation and Control*. Oxford: Oxford University Press.

Ball, G., Ling, D., Kurlander, D., Miller, D., Pugh, D., Skelly, T., Stankosky, A., Thiel, D., Van Dantzich, M. and T. Wax (1997). Lifelike computer characters: the persona project at Microsoft Research. In J. M. Bradshaw (ed.) *Software Agents*, Cambridge, MA: MIT Press.

Beattie, G. W. (1981). Sequential temporal patterns of speech and gaze in dialogue. In T.A. Sebeok and J. Umiker-Sebeok (eds.), *Nonverbal Communication, Interaction, and Gesture: Selections from Semiotica*. The Hague: Mouton.

Bavelas, J. Chovil, N., Lawrie, D. and Wade, A. (1992). Interactive Gestures. *Discourse Processes*, 15:469-489.

Beskow, J. and McGlashan, S. (1997). Olga—a conversational agent with gestures. Proceedings of Workshop on Animated Interface Agents. IJCAI-97 (August, Nagoya, Japan).

Bolt, R.A. (1980). Put-that-there: voice and gesture at the graphics interface. *Computer Graphics*, 14(3): 262-270.

Bregler, C., Hild, H., Manke, S., and Waibel, A. (1993). Improving connected letter recognition by lipreading. *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing* (IEEE-ICASSP) (Minneapolis, MN).

Brennan, S. and Hulteen, E. (1995). Interaction and feedback in a spoken language system: a theoretical framework. *Knowledge-Based Systems*, 8(2-3).

Cassell, J. and Prevost, S. (in preparation). Embodied natural language generation: a framework for generating speech and gesture.

Cassell, J. and Thórisson, K. (forthcoming). The power of a nod and a glance: envelope vs. emotional feedback in animated conversational agents. *Applied Artificial Intelligence.*

Cassell, J., McNeill, D. and McCullough, K.E. (1998). Speech-gesture mismatches: evidence for one underlying representation of linguistic and nonlinguistic information. *Pragmatics and Cognition*., 6(2).

Cassell, J., Pelachaud, C., Badler, N.I., Steedman, M., Achorn, B., Beckett, T., Douville, B., Prevost, S. and Stone, M. (1994). Animated conversation: rule-based generation of facial display, gesture and spoken intonation for multiple conversational agents. *Computer Graphics (SIGGRAPH proceedings)*. 28(4): 413-420.

Cassell, J., Torres, O. and S. Prevost (in press). Turn taking vs. Discourse structure: how best to model multimodal conversation. In Wilks (ed.), *Machine Conversations*. The Hague: Kluwer.

Condon, W.S. and Osgton, W.D. (1971). Speech and body motion synchrony of the speaker-hearer. In D.H. Horton and J.J. Jenkins (eds.), *The Perception of Language*, pp. 150-184. Academic Press.

Dittman, A.T. (1974). The body movement-speech rhythm relationship as a cue to speech encoding. In S. Weitz (Ed.), *Nonverbal Communication*. New York: Oxford University Press.

Duncan, S. (1974). Some signals and rules for taking speaking turns in conversations. In S. Weitz (ed.), *Nonverbal Communication*. New York: Oxford University Press.

Efron, D. (1941). *Gesture and Environment*.. New York: King's Crown Press.

Ekman, P. (1979). About brows: emotional and conversational signals. In M. von Cranach, K. Foppa, W. Lepenies, and D. Ploog (eds.), *Human Ethology: Claims and Limits of a New Discipline*, pp. 169-249. New York: Cambridge University Press.

Ekman, P. and Friesen, W.V. (1984). *Unmasking the Face*. Palo Alto: Consulting Psychologists Press.

Ekman, P. and Friesen, W. (1969). The repertoire of nonverbal behavioral categories—origins, usage, and coding. *Semiotica*, 1: 49-98.

Elhadad, M., McKeown, K. and Robin, J. (to appear). Floating constraints in lexical choice. *Computational Linguistics*.

Elliott, C. 1997. I picked up Catapia and other stories: a mulitmodal approach to expressivity for "emotionally intelligent" agents. Proceedings of the First International Conference on Autonomous Agents, pp. 451-457, February 5-8, Marina del Rey, California, USA.

Hajicova, E. and Sgall, P. (1987). The ordering principle. *Journal of Pragmatics* 11: 435-454.

Halliday, M. (1967). *Intonation and Grammar in British English.* The Hague: Mouton.

Hinrichs, E. and Polanyi, L. (1986). Pointing the way: a unified treatment of referential gesture in interactive contexts. In A. Farley, P. Farley and K.E. McCullough (eds.), *Proceedings of the Parasession of the Chicago Linguistics Society Annual Meetings (Pragmatics and Grammatical Theory).* Chicago: Chicago Linguistics Society.

Hirschberg, J. (this volume) Intonation in Spoken Dialogue.

Iverson, J. and Goldin-Meadow, S. (1996). Gestures in blind children. Manuscript, Department of Psychology, University of Chicago.

Johnson, M., Bransford, J., Solomon, S. (1973). Memory for tacit implications of sentences. *Journal of Experimental Psychology*, 98(1): 203-205.

Johnston, M., Cohen, P. R., McGee, D., Pittman, J., Oviatt, S. L., and Smith, I. (1997). Unification-based multimodal integration. *Proceedings of the Annual Meeting of the Association for Computational Linguistics*. (ACL-97/EACL-97), July, Madrid, Spain.

Kendon, A. (1993). Gestures as illocutionary and discourse structure markers in southern Italian conversation. *Proceedings of the Linguistic Society of America Symposium on Gesture in the Context of Talk.*

Kendon, A. (1980). Gesticulation and speech: two aspects of the process. In M.R. Key (ed.), *The Relation Between Verbal and Nonverbal Communication*. The Hague: Mouton.

Kendon, A. (1974). Movement coordination in social interaction: some examples described. In S. Weitz (ed.), *Nonverbal Communication*. New York: Oxford University Press.

Kendon, A. (1972). Some relationships between body motion and speech. In A.W. Siegman and B. Pope (eds.), *Studies in Dyadic Communication*. New York: Pergamon Press.

Koda, T. and Maes, P. (1996). Agents with faces: the effects of personification of agents. Proceedings of Human-Computer Interaction '96, August, London, UK.

Krauss, R., Morrel-Samuels, P. and Colasante, C. (1991). Do conversational hand gestures communicate? *Journal of Personality and Social Psychology*, 61(5): 743-754.

Lester, J.C., Voerman, J.L., Towns S.G., and Callaway, C.B. (forthcoming) Deictic Believability: Coordinating Gesture, Locomotion, and Speech in Lifelike Pedagogical Agents. *Applied Artificial Intelligence.*

Levelt, W. (1989). *Speaking: From Intention to Articulation*. Cambridge, MA: MIT Press.

Liberman, M. and Pierrehumbert, J. (1984). Intonational invariance under changes in pitch range and length. In M. Aronoff and R.T. Oehrle (eds.), *Language Sound Structure: Studies in Phonology Presented to Morris Halle by His Teacher and Students*, pp. 157-233. Cambridge, MA: MIT Press.

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

McNeill, D. (forthcoming) Models of speaking (to their amazement) meet speech-synchronized gestures. In D. McNeill (ed.) *Language and Gesture: Window into Thought and Action*. Hillsdale, NJ: Lawrence Erlbaum Associates.

McNeill, D. (1992). *Hand and Mind: What Gestures Reveal about Thought.* Chicago: University of Chicago Press.

Nagao, K. and Takeuchi, A. (1994). Social interaction: multimodal conversation with social agents. Proceedings of the 12th National Conference on Artificial Intelligence (AAAI-94), vol. 1, pp. 22-28. (August 1-4, Seattle, Washington, USA).

Noma, T. and Badler, N. (1997). A virtual human presenter. Proceedings of Workshop on Animated Interface Agents. IJCAI-97 (August, Nagoya, Japan).

Oviatt, S. L. (1995). Predicting spoken disfluencies during human-computer interaction. *Computer Speech and Language*, 9(1): 19-35.

Pelachaud, C., Badler, N. and Steedman, M. (1996). Generating facial expressions for speech. *Cognitive Science*, 20(1): 1-46.

Perlin, K. and Goldberg, A. (1996). Improv: a system for interactive actors in virtual worlds. *Proceedings of SIGGRAPH 96*, pp. 205-216. (New Orleans, LA, USA).

Power, R. (1977). The organisation of purposeful dialogues. *Linguistics*, 17: 107-152.

Prevost, S. (1996). An information structural approach to monologue generation. *Proceedings of the 34th Annual Meeting of the Association for Computational Linguistics* (June, Santa Cruz, California, USA).

Prevost, S. and Steedman, M. (1994). Specifying intonation from context for speech synthesis. *Speech Communication,* 15: 139-153.

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

Rimé, B. (1982). The elimination of visible behavior from social interactions: effects of verbal, nonverbal and interpersonal variables. *European Journal of Social Psychology,* 12: 113-129.

Rogers, W.T. (1978). The contribution of kinesic illustrators toward the comprehension of verbal behavior within utterances. *Human Communication Research*, 5: 54-62.

Scherer, K.R. (1980). The functions of nonverbal signs in conversation. In and R.N. St. Clair and H. Giles (eds.), *The Social and Psychological Contexts of Language*, pp. 225-243. Hillsdale, N.J.: Lawrence Erlbaum Associates.

Scoble, J. (1993). Stuttering blocks the flow of speech and gesture: the speech-gesture relationship in chronic stutters. M.A. thesis, McGill University.

Steedman, M. (1991) Structure and intonation. *Language*: 67(2): 260-296.

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

Thórisson, K. R. (forthcoming). A mind model of multimodal communicative creatures and humanoids. *Applied Artificial Intelligence.*

Thórisson, K. R. (1994). Face-to-face communication with computer agents. AAAI Spring Symposium on Believable Agents Working Notes, pp. 86-90 (March 19-20, Stanford University, California).

Trevarthen, C. (1986). Sharing makes sense: intersubjectivity and the making of an infant's meaning. In R. Steele and T. Threadgold (eds.), *Language Topics: Essays in Honour of M. Halliday*, vol. 1, pp. 177-200. Amsterdam: J. Benjamins.

Vilhjalmsson, H. and Cassell, J. (1998). BodyChat: autonomous communicative behaviors in avatars. *Proceedings of ACM International Conference on Autonomous Agents.*

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

Whittaker, S. and O'Conaill, B. (1997). The role of vision in face-to-face and mediated communication. In K.E. Finn, A.J. Sellen, and S.B. Wilbur (eds.), *Video-Mediated Communication* , pp. 23-49. Hillsdale, NJ: Lawrence Erlbaum Associates.

Wilson, A., Bobick, A. and Cassell, J. (1996). Recovering the temporal structure of natural gesture. Proceedings of the Second International Conference on Automatic Face and Gesture Recognition.