Nudge Nudge Wink Wink: Elements of Face-to-Face Conversation for Embodied Conversational Agents

Justine Cassell

It will not be possible to apply exactly the same teaching process to the machine as to a normal child. It will not, for instance, be provided with legs, so that it could not be asked to go out and fill the coal scuttle. Possibly it might not have eyes. But however well these deficiencies might be overcome by clever engineering, one could not send the creature to school without the other children making excessive fun of it.
—Alan Turing, "Computing Machinery and Intelligence," 1950

The story of the automaton had struck deep root into their souls and, in fact, a pernicious mistrust of human figures in general had begun to creep in. Many lovers, to be quite convinced that they were not enamoured of wooden dolls, would request their mistresses to sing and dance a little out of time, to embroider and knit, and play with their lapdogs, while listening to reading, etc., and, above all, not merely to listen, but also sometimes to talk, in such a manner as presupposed actual thought and feeling.
—E. T. A. Hoffmann, “The Sandman,” 1817

1.1 Introduction

Only humans communicate using language and carry on conversations with one another. And the skills of conversation have developed in humans in such a way as to exploit all of the unique affordances of the human body. We make complex representational gestures with our prehensile hands, gaze away and towards one another out of the corners of our centrally set eyes, and use the pitch and melody of our voices to emphasize and clarify what we are saying.

Perhaps because conversation is so defining of humanness and human interaction, the metaphor of face-to-face conversation has been applied to human-computer interface design for quite some time. One of the early arguments for the utility of this metaphor gave a list of features of face-to-face conversation that could be applied fruitfully to human-computer interaction, including mixed initiative, nonverbal communication, sense of presence, rules for transfer of control (Nickerson 1976). However, although these features have gained widespread recognition, human–computer conversation has only recently become more than a metaphor. That is, just lately have designers taken the metaphor seriously enough to attempt to design computer interfaces that can hold up their end of the conversation, interfaces that have bodies and know how to use them for conversation, interfaces that realize conversational behaviors as a function of the demands of dialogue but also as a function of emotion, personality, and social convention. This book addresses the features of human-human conversation that are being implemented in this new genre of *embodied conversational agents*, and the models and functions of conversation that underlie the features.

One way to think about the problem that we face is to imagine that we succeed beyond our wildest dreams in building a computer that can carry on a face-to-face conversation with a human. Imagine, in fact, a *face-to-face Turing test*. That is, imagine a panel of judges challenged to determine which socialite was a real live young woman and which was an automaton (as in

Hoffmann's "The Sandman"). Or, rather, perhaps to judge which screen was a part of a video conferencing setup, displaying the human being filmed in another room, and which screen was displaying an autonomous embodied conversational agent running on a computer. In order to win at this Turing test, what underlying models of human conversation would we need to implement, and what surface behaviors would our embodied conversational agent need to display?

The chapters assembled here demonstrate the breadth of models and behaviors necessary to natural conversation. Four models, in particular, that inform the production of conversational behaviors are employed by the authors in this volume, and those are *emotion, personality*, *performatives,* and *conversational function*. All of these models are proposed as explanatory devices for the range of verbal and nonverbal behaviors seen in face-to-face conversation, and therefore implemented in embodied conversational agents (ECAs). In what follows, I examine these nonverbal behaviors in depth, as background to the underlying models presented in each chapter. But first, I describe briefly the nature of the models themselves.

Several authors address the need for models of personality in designing ECAs. In the work of André et al. (chap. 8), where two autonomous characters carry on a conversation that users watch, characters with personality make information easier to remember because the narration is more compelling. Their characters, therefore, need to be realized as distinguishable individuals with their own areas of expertise, interest profiles, personalities, and audiovisual appearance, taking into account their specific task in a given context. Each character displays a set of attitudes and actions, consistent over the course of the interaction, and revealed through the character's motions and conversations and interactions with the user and with other characters.

Ball and Breese (chap. 7) propose that the user's personality should also be *recognized*, so that the agent's personality can match that of the user. Churchill et al. (chap. 3) focus more generally on how to create *personable* characters. They suggest that success will be achieved when users can create thumbnail personality sketches of a character on the basis of an interaction. They also point out that personality should influence not just words and gestures, but also reactions to events, although those reactions should be tempered by the slight unpredictability that is characteristic of human personality.

What behaviors realize personality in embodied conversational agents? The authors in this book have relied on research on the cues that humans use to read personality in other humans: verbal style, physical appearance and nonverbal behaviors. These will be addressed further below. The importance of manipulating these behaviors correctly is demonstrated by Nass, Isbister, and Lee (chap. 13), who show that embodied conversational agents that present consistent personality cues are perceived as more useful.

Several authors also address the need for models of emotion that can inform conversational behavior. In the chapter by Badler et al. (chap. 9), the emotional profile of the ECA determines the style of carrying out actions that is adopted by that character. In the chapter by Lester et al. (chap. 5), where the ECA serves as tutor, the character exhibits emotional facial expressions and expressive gestures to advise, encourage, and empathize with students. These behaviors are generated from pedagogical speech acts, such as cause and effect, background information, assistance, rhetorical links, and congratulation, and their associated emotional intent, such as uncertainty, sadness, admiration, and so on.

Ball and Breese (chap. 7) describe not only generation of emotional responses in their ECA but also recognition of emotions on the part of the human user, using a Bayesian network approach. The underlying model of emotion that they implement is a simple one, but the emotion recognition that this model is capable of may be carried out strictly on the basis of observable features such as speech, gesture, and facial expression.

Like Lester, Poggi and Pelachaud (chap. 6) generate communicative behaviors on the basis of speech acts. However, Poggi and Pelachaud concentrate on one particular communicative behavior—facial expression—and one particular kind of speech act—performatives. Performatives are a key part of the communicative intent of a speaker, along with propositional and interactional acts. They can be defined as “the reason the speaker is communicating a particular thing—what goal the speaker has in mind,” and they include acts such as “wishing, informing, threatening.” Because Poggi and Pelachaud generate directly from this aspect of communicative intention, they can be said to be engaging not in speech to text but, on the contrary, in *meaning to face*.

Many of the authors in this volume discuss conversational function as separate from speech acts, emotion, and personality. Cassell et al. (chap. 2) propose a model of conversational function. In general terms, all conversational behaviors in the FMTB conversational model must support conversational functions, and any conversational action in any modality may convey several communicative goals. In this framework, four features of conversation are proposed as key to the design of embodied conversational agents.

• the distinction between propositional and interactional functions of conversation
• the use of several conversational modalities, such as speech, hand gestures, facial expression
• the importance of timing among conversational behaviors (and the increasing co-temporality or synchrony among conversational participants)

• the distinction between conversational behaviors (such as eyebrow raises) and conversational functions (such as turn taking)

All of the models described so far are proposed as ways of predicting *conversational behaviors and actions*. That is, each model is a way of *realizing* a set of conversational surface behaviors in a principled way. In what follows, we turn to those conversational behaviors and actions. We concentrate on the nonverbal behaviors, which are what distinguish embodied conversational agents from more traditional dialogue systems (for a good overview of the issues concerning speech and intonation in conversational interfaces and dialogue systems, see Luperfoy n.d.). In particular, we focus here on hand gesture and facial displays[1] and ignore other aspects of nonverbal behavior (such as posture, for example).

1.2 Overview of Nonverbal Behaviors

What nonverbal behaviors, then, do we find in human-human conversation? 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, McNeill, and McCullough 1999), as well as to follow the conversational process (Bavelas et al. 1995). In ECAs, then, gestures can be realized as a function of models of propositional and interactional content. Likewise, faces change expressions continuously, and many of these changes are synchronized to what is going on in concurrent conversation (see Poggi and

Pelachaud, chap. 6; Pelachaud, Badler, and Steedman 1996). Facial displays are linked to all of the underlying models mentioned above and described in this book. That is, *facial displays* can be realized from the interactional function of speech (raising eyebrows to indicate attention to the other's speech), emotion (wrinkling one's eyebrows with worry), personality (pouting all the time), performatives (eyes wide while imploring), and other behavioral variables (Picard 1998). Facial displays can replace sequences of words ("she was dressed [wrinkle nose, stick out tongue][2]") as well as accompany them (Ekman 1979), and they can 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 and Osgton 1971; Kendon 1972). *Eye gaze* is also an important feature of nonverbal conversational behavior. Its main functions are (1) 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 he or she follows), (2) 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 (3) to indicate personality characteristics (staring at a person to show that one won't back down) (Beattie 1981; Duncan 1974).

Although many kinds of gestures and a wide variety of facial displays exist, the computer science community until very recently has for the most part only attempted to integrate one kind of gesture and one kind of facial display into human-computer interface systems-that is, *emblematic* gestures (e.g., the "thumbs up" gesture, or putting one's palm out to mean "stop"), which are employed in the absence of speech, and *emotional* facial displays (e.g., smiles, frowns, looks of puzzlement). But in building embodied conversational agents, we wish to exploit the power of gestures and facial displays that function in conjunction with speech.

For the construction of embodied conversational agents, then, there are types of gestures and facial displays that can serve key roles. In natural human conversation, both facial displays and gesture add redundancy when the speech situation is noisy, give the listener cues about where in the conversation one is, and add information that is not conveyed by accompanying speech. For these reasons, facial display, gesture, and speech can profitably work together in embodied conversational agents. Thus, in the remainder of this chapter, I will introduce those nonverbal behaviors that are integrated with one another, with the underlying structure of discourse and with models of emotion and personality

Let's look at how humans use their hands and faces. In figure X.1, 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.

Figure 1.1
Hawley talking about mosaic tiles.

As is his wont, in the picture, 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 nonverbal behavior than he is. 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.

Gestures and facial displays such as those demonstrated by Mike Hawley can be implemented in ECAs 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.

1.3 Kinds of Gesture

1.3.1 Emblems

When we reflect on what kinds of gestures we have seen in our environment, we often come up with a 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 and Friesen 1969). For example, the American "V for victory" gesture can be made either with the palm *or* the back of the hand toward the listener. In Britain, however, a "V" gesture made with the back of the hand toward 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 percent of the gestures produced by speakers. Despite the paucity of emblematic gestures in everyday communication, it was uniquely gestures such as these that interested interface designers at one point. That is, computer vision systems known as “gestural interfaces” attempted 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 untenable as a gestural. And in terms of *tokens*, we simply don’t seem to make that many emblematic gestures on a daily basis. In ECAs, 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.

1.3.2 Propositional Gestures

Another conscious gesture that has been the object of some study in the interface community is the so-called propositional gesture (Hinrichs and 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 placeholder 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.

1.3.3 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, Bransford, and Solomon 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, and Colasante 1991), even though we attend to the information that they convey (Cassell, McNeill, and McCullough 1999).

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 toward 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, by 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 nonphysical 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 toward 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 batonlike 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).

These gesture types may be produced in a different manner according to the emotional state of the speaker (Badler et al., chap. 9; Elliott 1997). Or they may differ as a function of personality (André et al., chap. 8; Churchill et al., chap 3; Nass, Isbister, and Lee, chap. 13). Their content, however, is predicted by the communicative goals of the speaker, both propositional and interactional (Cassell et al, chap. 2). The fact that they convey information that is not conveyed by speech, and that they convey it in a certain manner, gives the impression of cognitive activity over and above that required for the production of speech. That is, they give the impression of a *mind*, and therefore, when produced by embodied conversational agents, they may enhance the believability of the interactive system. But exploiting this property in the construction of ECAs requires an understanding of the *integration* of gesture with speech. This is what we turn to next.

1.4 Integration of Gesture with Spoken Language

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

1.4.1 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 versus nonrepresentational. 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, and 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." This issue of timing is a difficult one to resolve in ECAs, as discussed by Lester et al. (chap. 5), Rickel et al. (chap. 4) and Cassell et al. (chap. 2).

At the level of the turn, the hands being in motion is one of the most robust cues to turn taking (Cassell et al., chap. 2; 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.

1.4.2 Semantic Integration

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

In the previous section, I wrote that gesture is co-temporaneous 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,” and “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 n.d.). 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 versus 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 said “Road Runner just [goes]” and with one index finger extended made 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). This aspect of the relationship between speech and gesture is an ongoing research issue in psycholinguistics but has begun to be implemented in ECAs (Cassell and Stone 1999).

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 and 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, or 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.

1.4.3 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 where one finds gesture in discourse. 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 that 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 (Hajicova and Sgall 1987; Halliday 1967), one can 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 on what question was asked, the contribution of the current answer will be different.[3]

In English, intonation serves an important role in marking information as rhematic and as contrastive. 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 boldface type):

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

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” that 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 and Prevost n.d.). In this sense, intonation and gesture serve similar roles in the discourse. Intonational contours also time the occurrence of gesture (Cassell and Prevost n.d.). 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.

• Second, 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. Although we know that intonation and gesture function in similar ways, and are synchronized to one another, how to implement this property in ECAs is still an unsolved issue, due in part to the difficulty of reconciling the demands of graphics and speech synthesis software.

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

1.4.4 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, speakers from some language communities demonstrate a greater number of gestures per utterance than others. This phenomenon 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 gesturing. 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. That is, all except for one American man who made one single gesture during his entire twenty-minute narration, a gesture that 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 toward the listener or toward themselves. For British speakers, the “V for victory” with the palm oriented toward 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 percent of the gestures found in everyday conversation. The four 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. Nevertheless, the ways in which hand gestures and other nonverbal behaviors are produced do certainly differ from culture to culture. And, as Nass, Isbister, and Lee (chap. 13) show, cultural identification in

ECAs is important to users' estimation of their abilities. This topic remains, then, an important one for future research.

1.5 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 and 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 certain performatives, such as "implore").

In addition to characterizing facial displays by the muscles or part of the body in play, or the amount of time that they last, we can also characterize them by their function in a conversation. Some facial displays have a phonological function—for instance, lip shapes that change with the phonemes uttered. It has been shown that such lip shapes can significantly improve the facility with which people understand "talking heads" (Bregler et al. 1993; Massaro et al., chap. 10). Some facial displays fulfill a semantic function, for example, nodding rather than saying "yes." 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, or a glance at the other person when one is finished speaking. 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).

1.5.1 Cultural Differences

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 facial behaviors whose description has more universal validity.

1.6 Integration of Verbal Displays with Spoken Language

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

1.6.1 Temporal Synchronization

Synchrony occurs at all levels of speech: phonemic segment, word, phrase or long utterance. Different facial motions are isomorphic to these groups (Condon and 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 he or she is going to say.

1.6.2 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), while 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 and Cook 1976; Chovil 1992; Collier 1985). The following describes where behaviors in each of these four categories occur, and how they function.

• *Planning:* 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 nonverbal 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 most 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, and Prevost n.d.). 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 toward 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 nonverbal behaviors fill the function that Brennan and Hulteen (1995) suggest is needed for more robust speech interfaces.

1.7 Another Example

Let us now end with another example from life. This excerpt from a human speaking will show the rules we have just discussed, in action.

Figure 1.2
". . . will interact."

Figure 1.2 shows Seymour Papert talking about embedding computing in everyday objects and toys. He breathes in, looks up to the right, then turns toward 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 toward the audience, raises his hands to chest level, and points with each hand toward 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 toward 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.

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 similar. For example, although nobody else is taking a turn, Papert still gazes away before taking the turn and gazes toward 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.

1.8 Conclusion

One of the motivations for embodied conversational agents—as for dialogue systems before them—comes from increasing computational capacity in many objects and environments outside the desktop computer—smart rooms and intelligent toys—in environments as diverse as a military battlefield or a children's museum, and for users as different from one another as we can imagine. It is in part for this reason that we continue to pursue the dream of computers without keyboards, which can accept natural untrained input and respond in kind. In situations such as these, we will need natural conversation, multiple modalities, and well-developed characters to interact with.

We still cannot build an embodied conversational agent with anything like the conversational skills of Mike Hawley or Seymour Papert. Our models of emotion, of personality, of conversation are still rudimentary. And the number of conversational behaviors that we can realize in real time using animated bodies is still extremely limited. But, as we begin to understand the abilities that underlie human conversation, and to appreciate the behaviors that make up human conversation, we approach the day when a face-to-face Turing test will become imaginable.[6]

Notes

Research leading to the preparation of this chapter was supported by the National Science Foundation (award IIS-9618939), AT&T, Deutsche Telekom, and the other generous sponsors of the MIT Media Lab. Thanks to Andrew Donnelly for so ably handling the administrative aspects of this project, to all of the students of the Gesture and Narrative Language research group for dedication above and beyond the call of duty to this book as to every other project we undertake, and to David Mindell and Cathy O'Connor for providing intellectual and physical contexts for writing. Finally, profound and heartfelt thanks to David McNeill for introducing me to this field, sharing his passion about the topic, and teaching me so much about scholarship and academic community.

1. Following Takeuchi and 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.

2. Square brackets indicate the extent of speech that is accompanied by a nonverbal behavior.

3. 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 that would, in any case, be beyond the scope of this chapter.

4. We call these behaviors "envelope" to convey the fact that they concern the outer envelope of communication, rather than its contents (Cassell and Thórisson 1999). A similar distinction between content and envelope is made by Takeuchi and 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).

6. Parts of this chapter appeared in somewhat different form in Cassell (2000).

References

Argyle, M. and M. Cook. 1976. Gaze and mutual gaze. Cambridge: Cambridge University Press.

Bavelas, J., N. Chovil, L. Coates, and L. Roe. 1995. Gestures specialized for dialogue. *Personality and Social Psychology Bulletin* 21(4):394–405.

Bavelas, J. N. Chovil, D. Lawrie, and A. Wade. 1992. Interactive gestures. *Discourse Processes* 15:469–489.

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*, 298–320. The Hague: Mouton.

Bolt, R. A. 1980. Put-That-There: Voice and gesture at the graphics interface. *Computer Graphics* 14(3):262–270.

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

Brennan, S., and E. Hulteen. 1995. Interaction and feedback in a spoken language system: A theoretical framework. *Knowledge-Based Systems* 8(2–3):143–151.

Cassell, J. 2000. More than just another pretty face: Embodied conversational interface agents. C*ommunications of the ACM***.** Forthcoming.

Cassell, J., and Prevost, S. N.d. Embodied natural language generation: A framework for generating speech and gesture. Forthcoming.

Cassell, J., and M. Stone. 1999. Living hand to mouth: Theories of speech and gesture in interactive systems. In *Proceedings of the AAAI Fall Symposium on Psychological Models of Communication in Collaborative Systems* (Cape Cod, Mass.), 34–42**.**

Cassell, J., and K. Thórisson. 1999. The power of a nod and glance: Envelope vs. emotional feedback in animated conversational agents. *Journal of Applied Artificial Intelligence* 13(3):519–538.

Cassell, J., D. McNeill, and K. E. McCullough. 1999. Speech-gesture mismatches: evidence for one underlying representation of linguistic and nonlinguistic information. *Pragmatics and Cognition* 7(1):1–34.

Cassell, J., C. Pelachaud, N. Badler, M. Steedman, B. Achorn, T. Becket, B. Douville, S. Prevost, and M. Stone. 1994. Animated conversation: Rule-based generation of facial expression, gesture and spoken intonation for multiple conversational agents. *Computer Graphics SIGGRAPH Proceedings 1994*, 413–420. New York: ACM SIGGRAPH.

Cassell, J., Torres, O. and S. Prevost. N.d. Turn taking vs. discourse structure: How best to model multimodal conversation. In Y. Wilks, ed., *Machine conversations*. The Hague: Kluwer. Forthcoming.

Chovil, N. 1992. Discourse-oriented facial displays in conversation. *Research on Language and Social Interaction* 25:163–194.

Collier, G. 1985. Emotional expression. Hillsdale, N.J.: Lawrence Erlbaum Associates.

Condon, W. S,. and W. D. Osgton. 1971. Speech and body motion synchrony of the speaker-hearer. In D. H. Horton and J. J. Jenkins, eds., *The perception of language*, 150–184. New York: 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*, 169–249. New York: Cambridge University Press.

Ekman, P., and W. Friesen. 1969. The repertoire of nonverbal behavioral categories—Origins, usage, and coding. *Semiotica* 1:49–98.

———. 1984. *Unmasking the face*. Palo Alto, Calif.: Consulting Psychologists Press.

Elliott, C. 1997. I picked up Catapia and other stories: A multimodal approach to expressivity for "emotionally intelligent" agents. In *Proceedings of the First International Conference on Autonomous Agents* (Marina del Rey, Calif.), 451–457.

Hajicova, E., and P. Sgall. 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 L. Polanyi. 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.

Iverson, J., and S. Goldin-Meadow. 1996. Gestures in blind children. Unpublished manuscript, Department of Psychology, University of Chicago.

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

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.

———. 1974. Movement coordination in social interaction: some examples described. In S. Weitz, ed., *Nonverbal communication*. New York: Oxford University Press.

———. 1980. Gesticulation and speech: Two aspects of the process. In M. R. Key, ed., *The relation between verbal and nonverbal communication*. The Hague: Mouton.

———. 1993. Gestures as illocutionary and discourse structure markers in southern Italian conversation. In *Proceedings of the Linguistic Society of America Symposium on Gesture in the Context of Talk*.

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

Levelt, W. 1989. *Speaking: From intention to articulation*. Cambridge, Mass.: The MIT Press.

Luperfoy, S. N.d. *Spoken dialogue systems*. Cambridge, Mass.: The MIT Press. Forthcoming.

McNeill, D. 1992. *Hand and mind: What gestures reveal about thought*. Chicago: University of Chicago Press.

———. N.d. Models of speaking (to their amazement) meet speech-synchronized gestures. In D. McNeill, ed., *Language and gesture: Window into thought and action*. Hillsdale, N.J.: Lawrence Erlbaum Associates. Forthcoming.

Nickerson, R. S. 1976. On conversational interaction with Computers. In R. M. Baecker and W. A. S. Buxton eds., *Readings in human computer interaction*, 681–693. Los Altos, Calif.: Morgan Kaufman.

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

Picard, R. 1998. *Affective computing*. Cambridge, Mass.: The MIT 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.

Scherer, K. R. 1980. The functions of nonverbal signs in conversation. In R. N. St. Clair and H. Giles, eds., *The social and psychological contexts of language*, 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, Department of Psychology, McGill University.

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

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