In *Computer Vision in Human-Machine Interaction*, R. Cipolla and A. Pentland, eds. Cambridge University Press, in press.

A FRAMEWORK FOR GESTURE GENERATION AND INTERPRETATION

Justine Cassell[1]
MIT Media Lab
20 Ames Street
Cambridge, MA 02139-4307
justine@media.mit.edu

Introduction

In this paper I describe ongoing research that seeks to provide a common framework for the generation and interpretation of spontaneous gesture in the context of speech. I present a testbed for this framework in the form of a program that generates speech, gesture, and facial expression from underlying rules specifying (a) what speech and gesture are generated on the basis of a given communicative intent, (b) how communicative intent is distributed across communicative modalities, and (c) where one can expect to find gestures with respect to the other communicative acts. Finally, I describe a system that has the capacity to interpret communicative facial, gestural, intonational, and verbal behaviors.

I am addressing in this paper one very particular use of the term "gesture" -- that is, hand gestures that co-occur with spoken language. Why such a narrow focus, given that so much of the work on gesture in the human-computer interface community has focused on gestures as their own language -- gestures that might replace the keyboard or mouse or speech as an direct command language? Because I don't believe that everyday human users have any more experience with, or natural affinity for, a "gestural language" than they have with DOS commands. We have plenty of experience with actions, and the manipulation of objects. But the type of gestures defined as (Väänänen & Böhm, 1993) "body movements which are used to convey some information from one person to another" are in fact primarily found in association with spoken language (90% of gestures are found in the context of speech according to McNeill, 1992). Thus if our goal is to get away from learned, pre-defined interaction techniques and create *natural* interfaces for normal human users, we should concentrate on the type of gestures that come naturally to normal humans.

Spontaneous (that is, unplanned, unselfconscious) gesture accompanies speech in most communicative situations, and in most cultures (despite the common belief to the contrary). People even gesture while they are speaking on the telephone (Rimé, 1982). We know that listeners attend to such kinds of gestures, and that they use gesture in these situations to form a mental representation of the communicative intent of the speaker.

What kinds of meanings are conveyed by gesture? How do listeners extract these meanings? Will it ever be possible to build computers that can extract the meanings from human gesture in such a way that the computers can understand natural human communication (including speech, gesture, intonation, facial expression, etc.)? When computers can interpret gestures, will they also be able to display them such that an autonomous communicating agent will act as the interlocutor in the computer? We imagine computers that communicate like we do, producing and understanding gesture, speech, intonation and facial expression, thereby taking seriously the currently popular metaphor of the computer as conversational partner.

Background

The Communicative Function of Hand Gestures

A growing body of evidence shows that people unwittingly produce gestures along with speech in many different communicative situations. These gestures have been shown to elaborate upon and enhance the content of accompanying speech (McNeill, 1992; Kendon, 1972), often giving clues to the underlying thematic organization of the discourse or the speaker's perspective on events. Gestures have also been shown to identify underlying reasoning processes that the speaker did not or could not articulate (Church and Goldin-Meadow, 1986).

Do gestures play any role in human-human communication? We know that gestures are still produced in situations where there is no listener, or the listener cannot see the speaker's hands (Rimé, 1982), although more gestures may be produced when an addressee is present (Cohen, 1977; Cohen & Harrison, 1973). Thus it appears that gestures must serve some function for the speaker, independent of any communicative intent. In addition, we know that information appears to be just about as effectively communicated in the absence of gesture -- on the telephone, or from behind a screen (Short, Williams & Christie, 1976; Williams, 1977), and thus gesture is not**essential** to the interpretation of speech. But it has been shown that when speech is ambiguous (Thompson & Massaro, 1986) or in a speech situation with some noise (Rogers, 1978), listeners do rely on gestural cues (and, the higher the noise-to-signal ratio, the more facilitation by gesture). And, when adults are asked to assess a child's knowledge, they are able to use information that is conveyed in the child's gesture (and that is not the same as that conveyed by the child's speech) to make that assessment (Goldin-Meadow, Wein & Chang, 1992; Alibali, Flevares & Goldin-Meadow, 1994). Finally, when people are exposed to gestures and speech that convey slightly different information, whether additive or contradictory, they seem to treat the information conveyed by gesture on an equal footing with that conveyed by speech, ultimately seeming to build one single representation out of information conveyed in the two modalities (Cassell, McNeill & McCullough, in press).

We suspect that hand gestures must be integral to communication when we examine their temporal relationship to other communicative devices. Hand gestures co-occur with their semantically parallel linguistic units, although in cases of hesitations, pauses or syntactically complex speech, it is the gesture which appears first (McNeill, 1992). At the most local level, 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). At the most global level, we find that the hands of the speaker come to rest at the end of a speaking turn, before the next speaker begins his/her turn. At the intermediate level, the phenomenon o co-articulation of gestural units is found, whereby gestures are performed rapidly, or their production is stretched out over time, so as to synchronize with preceeding and following gestures, and the speech these gestures accompany. An example of gestural co-articulation is the relationship between the two gestures in the phrases ``if you [write the check][2] then I'll [withdraw] the money for you?". After performing the gesture that depicts writing a check, the hands do not completely relax. Instead, the right hand remains in space and depicts withdrawing something (like a letter from a mailbox). Thus the occurrence of the word "withdraw", with its accompanying gesture, affected the occurrence of the gesture that accompanied "write the check".

The essential nature of gestures in the communicative situation is demonstrated by the extreme rarity of ‘gestural errors’. That is, although spoken language is commonly quite disfluent, full of false starts, hesitations, and speech errors, gestures virtually never portray anything but the speaker’s communicative intention. Speakers may *say* "left" and mean "right", but they will probably *point* towards the right. Listeners may correct speakers’ errors, on the basis of the speaker’s gestures (McNeill, 1992).

So, we can conclude that hand gestures serve a communicative function in face-to-face communication. Hand gestures are ubiquituous in face-to-face communication, and appear to be integral to the production and comprehension of language in face-to-face contexts. In fact, listeners take into account the information conveyed by gesture, even when this information is not redundant to the information conveyed in speech. We next ask whether all gestures function in the same way in communicative contexts, and which types of gestures are most common.

Kinds of Gestures Found in Human-Human Communication

Let us first look at the types of gestures that have been covered primarily in the extant literature on computer vision and human-computer interface, and then contrast that with the type of gestures that have been covered in the extant psychological or linguistic literature on gestures as an integral part of communication.

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 & 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. That is, 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. Why, then, do emblematic type gestures (such as putting up one hand to mean stop, or making a "thumbs up" gesture to mean that everything is okay) appear so often in the human-computer interface literature? I think it is because emblematic gestures are *consciously produced* and therefore easier to remember.

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, 1987). 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 received the attention that they have because they are *conscious witting* gestures available to our self-scrutiny.

We have, however, still ignored 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. This, I believe, is the reason that so many of these gestures have been ignored by the human-computer interface community: these gestures do not come immediately to mind when we reflect on the gestures we see around us. In case the fact of losing access to the form of a whole class of gestures seems odd, consider the analagous 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. 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, McNeill & McCullough, in press).

The spontaneous unplanned, more common gestures are of four types:

* *Iconic* gestures depict by the form of the gesture some feature of the action or event being described; such as the gesture of holding a tube with a handle that accompanies "Press the [handle of the caulking gun slowly as you move the nozzle across the window ledge that needs caulk]".

Iconic gestures may specify the manner in which an action is carried out, even if this information is not given in accompanying speech. For example, only in gesture does the narrator specify the essential information of how the handle of the caulk gun is to be manipulated.

Iconic gestures may also 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, one speaker at the Computer Vision Workshop 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 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 meeting went on and on" accompanied by a hand indicating rolling motion. There need not be a productive metaphor in the speech accompanying metaphoric gestures; sometimes the "metaphors" that are represented in gesture have become entirely conventionalized in the language. There does need to be a recognizable vehicle that mediates between the form of the gesture and the meaning of the speech it accompanies.

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. That is, for example, one speaker at the Computer Vision Workshop pointed to Sandy Pentland, and said "the work was based on work in [appearance-based constraints]". Here, Sandy is representing -- or standing in for -- work on appearance-based constraints.

* *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 tube of caulk that the narrator pointed to on the workbench, or non-physical discourse entities. An example of the latter might be pointing left and then right while saying "well, Roberto was looking at Pietro across the table. . .".

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. An example from the conference comes from one speaker who named other researcher's technique of modelling faces and then said "we [don't] do that; we [bung] them all together". During the word "don't", this speaker used both hands to demarcate or wave away the space to his right , and during "bung", he brought both hands together to demarcate a space directly in front of him. In this example, the speaker is positioning the techniques that he chose not to use to one side, and the techniques that he did use directly in front of him.

* *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. An example is "she talked first, I mean second" accompanied by a hand flicking down and then up.

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.

It is natural to wonder about the cultural specificity of these types of gestures. We often have the impression that Italians gesture more and differently than do British speakers. As far as the question of quantity is concerned, it is true 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 having 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.

As far as the nature of gesture is concerned, as mentioned above, emblems do vary widely from language to language community. The four gesture types described, however, have appeared in narrations in a variety of languages: English, French, Spanish, Tagalog, Swahili, Georgian, Chinese, etc. 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. 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.

The Problem

Figure 1: Complex Gesture Generation

In the figure to the left, Seymour Papert is talking about the advantages to embedding computing in everyday objects and toys. He 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; when he says "real behavior" he stresses the words. He also stresses "will interact" while looking towards the audience, raising his hands to chest level and pointing with each hand towards the other as if each hand is a device that is about to interact with the other.

The concept of interaction, of course, does not have a physical instantiation. The gesture produced by Papert is a metaphoric depiction of the notion of interaction. Note that this communicative performance comprises speech with intonation, facial movements (gaze) and hand gestures, all three behaviors in synchrony.

How could we possibly interpret such a gestural performance? That is, how could we have understood that two index fingers pointing towards the center were two lego robots preparing to dance with one another (the example that he gave of interaction)?

For, unlike language, gesture does not rely on a one-to-one mapping of form to meaning. That is, two fingers pointing towards the center may convey dancing robots at one point, and at another point in the very same discourse, that same gesture may indicate rolling up a carpet in a room. Of course, the fact that gesture is not a code, is what makes it a powerful index of human mental representation. Spoken languages are constrained by the nature of grammar, which is arbitrary and non-iconic (for the most part). Language is mediated by the ratified social code. Gesture, on the other hand, can play out in space what we imagine in our minds.

And yet, the psycholinguistic literature tells us that humans attend to gestures of just this sort, and interpret them in real-time. When speech and gesture do not convey the same information, listeners do their best to reconcile the two modalities. How is this done?

In what follows, I suggest that gesture interpretation may (a) be bootstrapped by its synchronization with intonation and the information structure of an utterance; (b) that gesture takes up the parts of a semantic representation not taken up by the speech, and that listeners may use gesture to fill in empty feature slots in a semantic frame; (c) our experience with objects in the world provides action frames that may be used to inform the semantic frames needed for gesture interpretation.

Previous Solutions

Implementations of Gesture in Multi-modal Systems

As I said above, vision systems have tended to concentrate on gestures *as* a language, rather than gesture as a part of a multi-modal communicative event. So-called multi-modal systems have to some extent suffered from the same problem, concentrating on gestures as replacements for words in the stream of meaning. Nonetheless, some of these systems have been quite successful in combining speech and gesture at the computer interface. One of the first such systems was *Put-That-There*, developed by Richard Bolt, Christopher Schmandt and their colleagues (Bolt, 1987; Bolt, 1980). *Put That There* used speech recognition and a six-degree-of-freedom space sensing device to gather input from a user's speech and the location of a cursor on a wall-sized display, allowing for simple deictic reference to visible entities. Recently, several systems have built on this attempt. (Koons et al, 1993) uses a two-dimensional map using spoken commands, deictic hand gestures, as well as deictic eye movement analysis (indicating where the user is looking on the display). In this system, nested frames are employed to gather and combine information from the different modalities. As in Put-that-There, speech drives the analysis of the gesture: if information is *missing* from speech (e.g. "delete that one"), then the system will search for the missing information in the gestures and/or gaze. Time stamps unite the actions in the different modalities into a coherent picture. (Wahlster, 1991) uses a similar method, also depending on the linguistic input to guide the interpretation of the other modalities. (Bolt & Herranz, 1992) describe a system that allows a user to manipulate graphics with semi-iconic gestures. (Bers, in press) developed a system that allows the user to combine speech and pantomimic gesture to direct a bee on how to move its wings -- the gestures are mapped onto the bee's body, making it move as prescribed by the user's pantomimic example. Bers developed a gesture segmentation scheme for his system that utilizes the kinetic energy of body part motion. Using a cutoff point and time stamping, motions can be selected that relate to the intended movement mentioned in speech. Sparrell (1993) used a scheme based on stop-motion analysis: whenever there is a significant stop or slowdown in the motion of the user's hand, then the preceding motion segment is grouped and analyzed for features such as finger posture and hand position. In all of these systems interpretation is not carried out until the user has finished the utterance.

Missing from these systems is a concept of non-verbal function with respect to discourse function. That is, in the systems reviewed thus far, there is no discourse structure over the sentence (no notion of "speaking turn" or "new information"). Therefore the role of gesture and facial expression cannot be analyzed at more than a sentence-constituent-replacement level. What is needed is a discourse structure that can take into account turn-taking, and the increasing accumulation of information over the course of a discourse. In the next sections I describe a discourse framework for the generation of multi-modal behaviors, and then how such a framework is being used to integrate and interpret multi-modal behaviors incrementally (that is, before the user has finished the utterance).

The Current Solution

A Discourse Framework for Gesture

What I am suggesting is that gesture fits into the entire context of communicative activities in particular rule-governed ways, and that understanding the interaction of those communicative modalities will help us build systems that can interpret and generate gesture. In particular, understanding the parallel and intersecting roles of information structure, intonation (the prosody or "melody" of language), and gesture, will give us a discourse framework for predicting where gestures will be found in the stream of speech.

Information Structure

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, 1987), we use the terms *theme* and *rheme* to denote two distinct information structural attributes of an utterance. The theme/rheme distinction is often referred to in the literature by the terms *topic/comment* or *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. The following examples illustrate the coupling of intonational "tunes" with themes and rhemes[3].

[Q:] Who withdrew fifty dollars?

[A:] (**George**) RHEME (withdrew fifty dollars) THEME

[Q:] What did George withdraw?

[A:] (George withdrew)THEME (**fifty dollars**)RHEME

If you speak these examples aloud to yourself, you will notice 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* . For the project at hand, an understanding of the relationship between information structure and intonation is crucial because informaion structure may predict in which utterances gestures occur, and intonation predicts the timing of gestures with respect to the utterance in which they occur.

It has been suggested that "intonation belongs more with gesture than with grammar" (Bolinger, 1983). Not only do intonation and hand and face gestures function in similar ways, they also stand in simil relationships to the semantic and information structures underlying spoken language. That is, we believe that the distribution of gestural units in the stream of speech is similar to the distribution of intonational units, in three 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") cooccurs with the pitch accent, or most effortful part of enunciation.

* Third, we hypothesize that one is most likely to find gestures co-occuring with the rhematic part of speech, just as we find particular intonational tunes co-occuring with the rhematic part of speech. We hypothesize this 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 convey additional content to speech, 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.

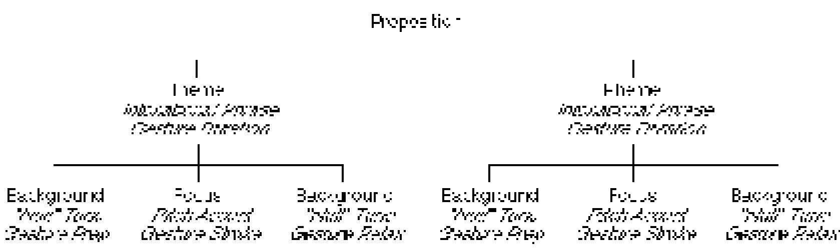

Figure 2: Discourse Framework

The Role of the Discourse Framework in Generation

Let’s turn to how this framework might allow us to automatically generate gestures along with speech and intonation. In the system "Animated Conversation" (Cassell*et al* , 1994) we automatically generated speech content, intonation, hand gestures (and facial expression, which we won’t address further here) on the basis of the information given above. In particular, gestures were generated along with words or phrases that the discourse model had marked as rhematic. The domain which we addressed was banking; that is, the conversation that was generated was between a bank teller and a customer desirous of withdrawing money. As far as*which kind* of gesture was generated in a given context, we relied on the taxonomy of gesture discussed above, and implemented it 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").

The timing of gestures was also implemented according to the psycholinguistic findings described above. Information about the duration of intonational phrases acquired in speech generation was then used to time gestures. If there was a non-beat gesture in an utterance, its preparation was set to begin at or 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 so as to coincide with the stressed syllable of the word that realizes the associated concept.

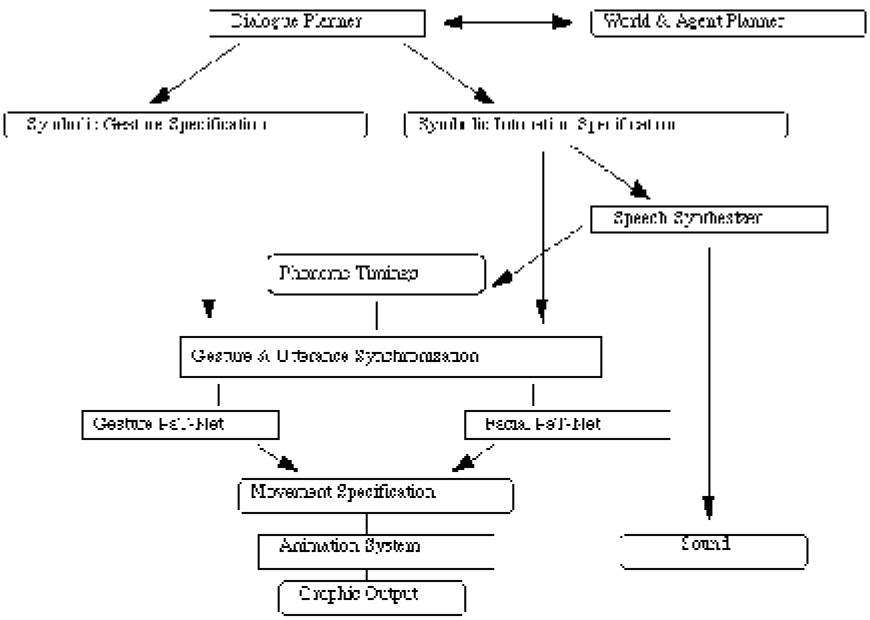

Figure 3: Animated Conversation System Architecture

All of the modalities were generated on an equal footing. That is, we based our system on the claim that gesture and speech are generated by humans at the same ‘computational stage’ (McNeill, 1992), and that gesture is not a *translation* of speech, but its equal partner in communication. Thus, the dialogue generation engine at the top of the system was 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.

This heuristic for generating verbal and nonverbal behaviors was successful to a certain extent. Gestures looked natural, and occured in natural looking places. They appeared timed correctly with respect to intonation. In Figure 4 and Figure 5 are reproduced excerpts from the conversation generated by the Animated Conversation system.

Figure 4: (a) "do you have a blank check?"; (b) "can you help me?"

Figure 4 (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 5: (a) "you can write the check"; (b) "I have eighty dollars"

Figure 5 (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 the generation of a beat gesture along with the phrase "yes, I have eighty dollars in my account".

However, although the distribution of gestures in the discourse, and the choice of type of gestures were generated automatically , we relied on a provisional and unsatisfactory method for generating the form of particular gestures. A gesture "dictionary" was accessed, which provided particular gesture forms for the domain of the dialogue.

A Semantic Framework for Gesture

The next step, then, is to further deconstruct gestures so as to understand how particular forms of gesture are generated in particular instances. Once again, we start by attempting to generate gesture, but we keep in mind the goal of a framework that will also allow interpretation of naturally occuring gesture. This goal will be partially addressed by the system described in the last section of the paper.

We start from the premise that we don't have access to a speaker's communicative intentions, and we don't always have access to the scene or events being described by a speaker, but we do have access to language -- to the common linguistic code that we share with our interlocutors. And, an essential part of the knowledge of language is the knowledge of lexical semantics.

Lexical Semantics

The lexicon is our mental vocabulary, the words that we know, in all of their forms (I write, you write, she writes) along with their meanings. When we speak, we choose words from the lexicon, make them fit the other words around them and then utter them. Semantics is the study of, or a theory of how the meanings of words are related to one another, and to our thinking processes, and how particular words are chosen. That is, the concept that underlies the verb "go" includes the concepts that underlie the verbs "walk", "run", "drive", etc. because those are different manners of going somewhere. Here we are particularly interested in the issue of*lexical choice* , or how one word is chosen over another. Why do we say "I *hightailed* it out of the room" rather than "I *left* the room"? Why do we choose to say "Sandy walked to the conference" one day, and another day "Sandy went to the conference on foot"? In the first sentence the manner of locomotion is conveyed in the verb, and in the second sentence, the manner of locomotion is conveyed in the prepositional phrase.

A Lexical Semantics that includes Gesture

What does any of this have to do with gesture? Well, we have rejected the idea of a dictionary of gestures that speakers draw from to produce gestures and that listeners draw from for gesture interpretation because of evidence about the absence of a one-to-one mapping between form and meaning in everyday gesture. We know, however, that gestures do sustain a tight link with the semantics of speech. We described above the evidence on the non-overlapping and yet complementary nature of the information conveyed in speech and gesture. Here we hypothesize that, for material that has been chosen to be conveyed, the concepts are represented mentally in terms of a complex of semantic features, and some of those features are conveyed in speech, and some are conveyed in gesture.

We are currently implementing this model by building a dialogue generation system that annotates for theme and rheme (Prevost, 1996), and takes semantic features into account when distributing information across speech and gesture. This resolves some of the problems seen in'Animated Conversation'.

Thus, in some cases one would hear "Hannes walked to the store", while in other cases one would hear "Hannes went to the store" and see the following gesture indicating the manner of locomotion.

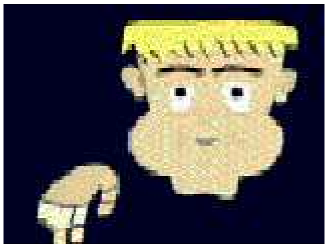

Figure 6: Distribution of Semantic Features in Speech and Gesture: "Hannes went to the store"

The examples given thus far account for the choice of gestures that accompany verbs of motion, primarily. These gestures mostly represent the viewpoint of an observer on the action. That is, the speaker's hand as a whole represents a characters feet, or body (such as in Figure 6). Many gestures, however, are more pantomimic and involve the representation of an action where the speaker's hands represent somebody's hands. For example, a speaker might say "and then I saw him pound in the nail" and with her hands represent somebody holding a hammer, and hammering. In this case, although the sentence refers to the speaker herself ("I") as well as some third person, the gesture represents action on the part of the third person. Gestures such as these involve representation and knowledge of action. In order for such gestures to be accounted for in a theory of lexical choice, the semantics must be of a form that allows knowledge of the world.

The Link between Gesture and Action

If the form of gestures is partially determined by semantic features, how are those semantic features represented mentally such that they can be instantiated in gesture? I believe that some of the semantic features that are represented in iconic gestures are acquired through action schemata. In order to illustrate this process, let’s look at the development of gestural sematic features from a series of actions.

The Acquisition of Gestural Features in Real Time

Imagine the following scenario. Lucinda is watching "This Old House", a television show about do-it-yourself home renovation[4]. During the introductory credits we see the hero of the show repairing windows in a beautiful old home. During the show, the narrator, talking about weatherproofing a Victorian home, is describing the caulking gun. The narrator picks up a caulking gun from the table in front of him and introduces its use:

"This is a caulking gun, which one fills with tubes of caulk, and which is used to fill and waterproof exposed wood."

As the narrator speaks, he lifts the handle to show where the caulk tube is inserted, and lowers the handle to show how to extrude the caulk. He also points to a tube of caulk on the workbench. He then replaces the tool on the workbench, and continues his discussion of exposed wood by explaining how to waterproof window ledges. In this second part of the discussion, the narrator describes how to use the caulking gun. The narrator starts, however, by framing the relevance of his talk:

"In this [next part]A I’m going to tell you how to use a caulking gun to [prevent leakage]B through [storm windows]C and [wooden window ledges]D (. . .) Press the [handle of the caulking gun slowly as you move the nozzle across the window ledge that needs caulk]E".

The narrator makes the following gestures during this segment of talk:

A. The narrator opens his hands, with the palms facing one another and the fingers facing away from his body, and then moves his wrists so that his fingers are facing down -- as if he is delineating a box.

B, C, D. The narrator’s right hand is lax, but he flips it over and back so that the palm is facing first upward and then downward.

E. The narrator forms his left hand into a fist to represent the body of the caulking gun, holding it diagonally in front of his body, and uses his right hand to represent the hand of the user, pumping an imaginary handle up and down.

Lucinda goes to Home Depot to pick up materials to repair her leaking windows. She says to the salesperson "where do I find that . . . [gesture ‘E’] to fill the cracks in my window ledges?"

Lucinda has learned the concept of ‘caulking gun’ from a linguistic-action-gestural performance[5]. Two things should be noted about the performance that Lucinda observes:

* The description of the concept of caulking proceeds by first *demonstrating* caulking on a actual house (during the introductory credits), next *defining* caulking through a generic description of a caulking gun, and finally *describing* how to caulk. The passage is, therefore, from a particular present event of caulking the narrator’s house, to a generic event of caulking, to another particular future event of caulking the listener’s house.

* While the narrator is first defining the caulking gun in speech, he is adding non-redundant features to that definition with his actions. Only in his actions does he show the relationship between the caulking gun and the tube of caulk (e.g. that the tube is fit into the cradle of the caulking gun), and the manner of using the caulking gun (e.g. that pushing down the handle extrudes the caulk). Likewise, in the second part of the narrator’s description, when the describes how to caulk, the speech and gesture are non-redundant: only in gesture is the manner of movement demonstrated. By this second part of the performance, the hands have become a *symbol* of the object spoken of: an iconic representation of the tool and the home owner’s interaction with i[6]. Note that not just representational gestures are used here; three other qualitatively different gestures are also a part of this communicative act. These four gesture types enable the constuction of a sensorimotor schema for adults: a semantic frame or schema that encompasses the knowledge types that allow the listener to understand the talk about home repair, and will allow the listener to caulk her own house later on.

Thus, in the time of one communicative event, the context of talk has gone from being the world itself -- as populated by do-it-yourself-show hosts, caulking guns, and tubes of caulk -- to being a representation of the world. This representation by the speaker is what allows the listener to construct a representation of the activity herself: including new lexical items, images of tools and activities, and motor knowledge (knowledge of how-to). Thus, the background against which talk is being interpreted by the listener has gone from being the Victorian home in the country, to the studio of "This Old House", to an imaginary event of home renovation. Although the *focal event* (Goodwin & Duranti, 1992) has remained the same -- the use of a caulking gun -- the *background* has changed several times.

The relevance of such an action schema approach to gestural meaning for our concerns here is two-fold. First, as far as gesture generation is concerned, we might envisage the generation of task-oriented dialogues that share semantic features between the simulation of physical actions and the instruction talk (and gestures) that accompanies the actions. That is, the animated agents that we create may be able to follow instructions (Webber, 1994; Webber *et al* , 1995), and also use their knowledge of the world and objects around them to give instructions. Secondly, as far as interpretation is concerned, we can envisage a discourse understanding system that learns semantic information from task oriented talk and physical actions, and then goes on to apply that knowledge to the understanding of gestures in talk about the same domain.

So far, then, we have a discourse framework, and a semantic framework, as follows:

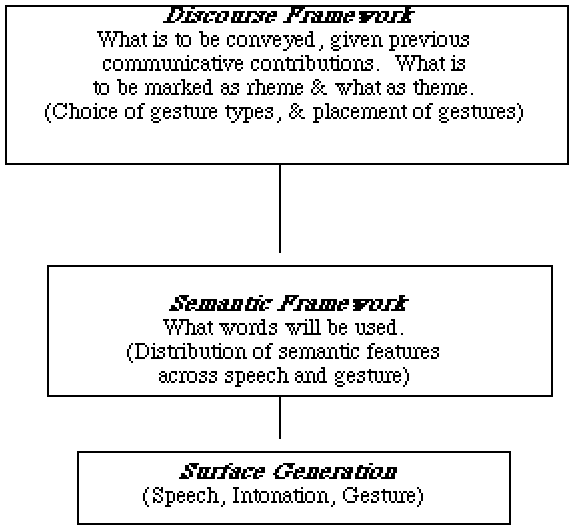

Figure 7: Model of Multi-Modal Generation

Applying the Solution to Interpretation

The framework above provides a general theory of the relationship between verbal and nonverbal behavior, but our implementations up until this point have been in the form of dialogue *generation* systems. In what follows, I turn to the issue of *interpreting* multi-modal behavior of the spontaneous unplanned sort that we have described.

I use the word "interpretation" rather than ‘recognition’ or ‘perception’ because we will beg the issue of vision by using input gathered through data gloves, a body tracker, a gaze tracker, and a microphone for speech.

Supposing a perceptual system, we still need an interpretation layer that can integrate information from gaze, hand gesture, intonation, and speech. Ymir is 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 (Thórisson, 1995, 1996). 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 accomodate two types of behavior. On the one hand, some communicative behavior controls the *envelope of communication*. For example, gaze is an indicator to the participants of a conversation for deciding 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, 1972). On the other hand, some communicative behavior controls the *content of communication*. For example, the content of speech, and the content of iconic gestures determine the direction that the conversation is taking. Ymir has layers dedicated to *reactive* behaviors such as gaze and other turn-taking signals and *reflective* behaviors such as speech and contentful gestures. 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.

Gandalf (Thórisson, *op cit*) is the first agent constructed in this architecture. It has been provided with the minimal behaviors necessary for face-to-face dialogue. It can understand limited utterances (currently using a grammar-based speech recognizer), intonation, body stance (oriented towards Gandalf or towards the task at hand), 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. 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. We are currently adding the ability to understand and generate additional behaviors.

Figure 8: Gandalf understands and responds to speech, gaze & hand gestures

The Ymir system, as noted above, relies on gestural input gathered by way of cybergloves and a body tracker. The hardware route was chosen because of the difficulty of using vision systems to recognize the handshape and consequently the meaning of gestures (due to occlusion, etc.). But not all discourse tasks require a recognition of gesture handshape. A *classification* of gestures into iconics, deictics and beats would also be helpful as a way of bootstrapping the understanding of the concurrent speech. In addition, if one could distinguish iconics from beats, then one could concentrate the vision resources of a system on recognizing the form of iconics, and simply acknowledge the presence of beats. To this end, I am currently exploring a more *perception-based* approach to the classification of gesture, in terms of temporal phase. I believe that gestures classified by temporal phase will turn out to correlate significantly with gestures classified by function, thus facilitating the use of computer vision to extract meaning from gesture. A preliminary attempt to implement this theory in a vision system has met with encouraging although limited results (Wilson, Bobick & Cassell, 1996).

Conclusions

In this paper, the characteristics of natural spontaneous human gestures were described, and the types of gestures that have served as the focus for many human-computer interface systems were contrasted with the types of gestures that are found in human-human communication. Although there are good reasons for having focused on the former type of gesture -- they are the types of gestures more accessible to reflection, and easier to treat as linguistic units -- it was argued that it is time to address in our human-computer interfaces the whole range of gesture types. To this end, a framework that allows for the understanding of gesture in the context of speech was presented. Several implementations of this framework were presented. The more advanced system deals with the generation of multi-modal communicative behaviors, but a new system provides a context for the interpretation of gestures in the context of speech. Although the systems presented are by no means complete, they encourage us to push forward in the use of natural spontaneous gesture and speech in our communication with computers.

References

Alibali, M.W., Flevares, L. & Goldin-Meadow, S. (1994). *Going beyond what children say to assess their knowledge.* Manuscript, Department of Psychology, University of Chicago.

Bers, J. (in press). A body model server for human motion capture and representation. *Presence: Teleoperators and Virtual Environments*, 5(4).

Bolinger, D. (1983). Intonation and gesture. *American Speech*, 58.2, 156-174.

Bolt, R.A. (1987). The integrated multi-modal interface. *Transactions of the Insititute of Electronics, Information and Communication Engineers (Japan)*, J79-D(11), 2017-2025.

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

Bolt, R.A. & Herranz, E. (1992). Two-handed gesture in multi-modal natural dialog. *Proceedings of OISI '92, Fifth Annual Symposium on User Interface Software and Technology*, Monterey, CA.

Cassell, J., McNeill, D. & McCullough, K.E. (in press). Speech-gesture mismatches: evidence for one underlying representation of linguistic & nonlinguistic information. *Cognition*.

Cassell, J., Pelachaud, C., Badler, N.I., Steedman, M., Achorn, B., Beckett, T., Douville, B., Prevost, S. & Stone, M. (1994a). Animated Conversation: Rule-based generation of facial expression, gesture and spoken intonation for multiple conversational agents. *Computer Graphics (SIGGRAPH proceedings)*.

Cassell, J., Steedman, M., Badler, N., Pelachaud, C., Stone, M., Douville, B., Prevost, S., & Achorn, B. (1994b). Modeling the interaction between gesture and speech. *Proceedings of the Cognitive Science Society Annual Conference*.

Cohen, A.A. (1977). The communicative functions of hand illustrators. *Journal of Communication*, 27(4), 54-63.

Cohen, A.A. & Harrison, R.P. (1973). Intentionality in the use of hand illustrators in face-to-face communication situations. *Journal of Personality and Social Psychology*, 28, 276-279.

Church, R.B. & Goldin-Meadow, S. (1986). The mismatch between gesture and speech as an index of transitional knowledge. *Cognition*, 23, 43-71.

Duncan, S. (1972) Some signals and rules for taking speaking turns in conversations. *Journal of Personality and Social Psychology*, 23 (2): 283-292.

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

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

Goldin-Meadow, S., Wein, D. & Chang, C. (1992). Assessing knowledge through gesture: Using children’s hands to read their minds. *Cognition and Instruction*, 9(3), 201-219.

Goodwin, C. & Duranti, A. (1992). Rethinking context: An introduction. *Rethinking Context*. Cambridge, England: Cambridge University Press.

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

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

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

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.* Mouton.

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

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

Krauss, R., Morrel-Samuels, P. & Colasante, C. (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, MA: MIT Press.

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

Prevost, S. (1996). "An information structural approach to monologue generation". Proceedings of the 34th annual meeting of the Association for Computational Linguistics (June, 1996; Santa Cruz).

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.

Short, J., Williams, E., & Christie, B. (1976). *The social psychology of telecommunications*. New York: Wiley.

Sparrell, C.J. (1993). Coverbal iconic gesture in human-computer interaction. Master's thesis, Massachusetts Institute of Technology. Cambridge, MA.

Thompson, L.A. & Massaro, D.W. (1986). Evaluation and integration of speech and pointing gestures during referential understanding. *Journal of Experimental Child Psychology*, 42, 144-168.

Thorisson, K. R. (1995). Multimodal Interaction with Humanoid Computer Characters. Conference on Lifelike Computer Characters, Snowbird, Utah, September 26-29, p. 45 (abstract).

Thorisson, K. R. (1996). Communicative Humanoids: A Computational Model of Psychosocial Dialogue Skills. Ph.D. Thesis, Media Arts and Sciences, M.I.T. Media Laboratory, unpublished.

Väänänen & Böhm, K. (1993). Gesture-driven interaction as a human factor in virtual environments-an approach with neural networks. In R.A. Earnshaw, M.A. Gigante & H. Jones (Eds.), *Virtual Reality Systems*. London: Academic Press Ltd.

Wahlster, W., Andre, E., Graf, W. & Rist, T. (1991). Designing illustrated texts. *Proceedings of the Fifth EACL:* 8-14.

Webber, B. (1994). Instruction Understanding for Human Figure Animation. Proc. 1994 AAAI Spring Symposium on Active Natural Language Processing. Stanford CA, March 1994.

Webber, B., Badler, N., DiEugenio, B., Geib, C., Levison, L., & Moore, M. (1995). Instructions, Intentions and Expectations. *Artificial Intelligence Journal* 73.

Williams, E. (1977). Experimental comparisons of face-to-face and mediated communication: A review. *Psychological Bulletin*, 84, 963-976.

Wilson, A., Bobick, A. & Cassell, J. (1996). Recovering the Temporal Structure of Natural Gesture. Submitted to the Second International Conference on Automatic Face and Gesture Recognition.