Multimodal Dialogue: Psychological and Interface Research

3.

In this chapter we will look at some of the psychological and computational research relevant to the task of building face-to-face interfaces. When explicitly applying the face-to-face metaphor to computer systems, 3 interdependent elements stand out:

- Dialogue structure. *The structure of human face-to-face dialogue is organized around the turn taking system. This system has the properties of requiring real-time responsiveness and concurrent input and output.*
- Multiple modes. *The inputs and outputs are multimodal, including speech, gesture and other visible behaviors.*
- Embodiment. *Face-to-face interaction requires participants that are embodied, which in turn gives meaning to their situated visual and auditory behavior.*

These will be used to focus the discussion in this chapter. In addition, an overarching theme is the notion of reciprocity in dialogue. Reciprocity is not only a major part of content coordination, as convincingly shown by numerous researchers [Clark & Brennan 1990, Goodwin 1986, Grosz & Sidner 1986, Kahneman 1973], but part of all elements of discourse. A major assumption in this work is that for the multimodal conversation metaphor to reach its full potential, we need to support the full feedback loop from user to machine and back, and address the metaphor's key elements on all levels. The following discussion will therefore necessarily be broad, covering first the structure of dialogue at all levels, as well as multiple modes and embodiment, and then go into implemented computer systems based on the multimodal metaphor.

3.1 Human Multimodal Communication

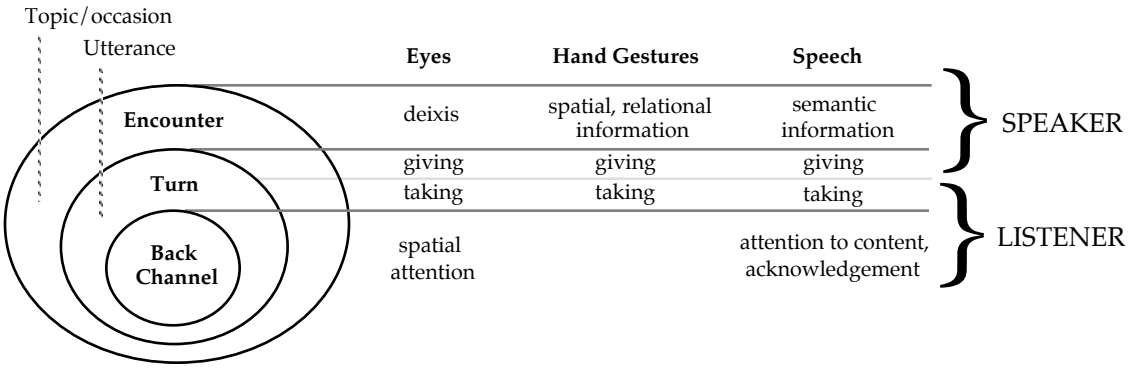
3.1.1 Dialogue Structure

Recent research in linguistics has indicated that in discourse, communicating parties strive to reach a common ground, a process that has been referred to as grounding [Clark 1992, Whittaker & Walker 1991, Clark & Brennan 1990]. The success of the grounding process depends on the successful support of dialogue by the common organizational principles of turn-taking, back-channel feedback and other multimodal communicative mechanisms [Sacks et al. 1974], as well as on focus of attention, indicated through gaze and spatial orientation of the interlocators, directing each other's attention with gaze and gestures [Clark & Brennan 1990, Goodwin 1986, Grosz & Sidner 1986, Kahneman 1973]. For the current purposes, we will adopt a 3-level hierarchical model of face-to-face interaction according to dependencies of its coordination constituents and the granularity of time. The highest level can be said to be the encounter. The encounter includes the whole interaction sequence that occurs when two or more people meet, including greetings and good byes [Schegloff & Sacks 1973], choice of topic, reason for the meeting, etc. Actions at this level happen at the slowest rate. The psychosocial actions in a conversation happen at the next two levels down, the first of which is the turn, the second being the back-channel. We shall now look at each in turn.

3.1.2 Turn Taking

When people communicate in face-to-face interaction they take turns speaking [Duncan 1972]. Goodwin [1981, p. 2] says about the turn:

FIGURE 3-1. The three main processes in face-to-face interaction can be thought of as hierarchically nested within each other (circles) according to their time span and time-criticality; the functional roles of speech, gesture and gaze in each conversational process are shown to the right.



> "In the abstract, the phenomenon of turn-taking seems quite easy to define. The talk of one party bounded by the talk of others constitutes a turn, with turn-taking being the process through which the party doing the talk of the moment is changed."[1]

The turn system's main function is to manage the sequential nature of talk. It organizes the information exchange between two (or more) communicating parties and ensures efficient transmission between them. The information can be constructed through speech, hand gestures, body language, gaze, facial expressions, or any combination thereof [Sacks 1992b, McNeill 1992, Goodwin 1981]. Turn-taking and back-channel feedback have both been shown to be important for conducting successful dialogue [Sacks et al. 1974, Nespolous & Lecours 1986]. Turn-taking is, for example, crucial in both negotiation and clarification [Whittaker et al. 1991, Whittaker & Stenton 1988, Sacks et al. 1974].

Sacks et al. [1974] put forth a model of turn taking that models the structure of human conversation as an emergent property of local decisions based on prediction by the participants. Because theirs is a thorough model, as psychological models go, and relates directly to psychosocial dialogue skills, I will briefly recap its main points. In their view, turn taking is locally managed and participant-administrated. Local management means that "all the operations [within the system] are 'local', i.e. directed to 'next turn' and 'next transition' on a turn-by-turn basis" [p. 725]. In this view, any pattern that arises out of interaction is "emergent"—i.e. results from the interaction of rules. They say further [p. 725-6] that

> "the turn-taking system is a local management system ... in the sense that it operates in such a way as to allow turn-size and turn-order to vary and be under local management, across variations in other parameters, while still achieving both the aim of all turn-taking systems—the organization of 'n at a time'—and the aim of all turn-taking organizations for speech-exchange systems—'one at a time while speaker change recurs'".

Party-administration refers to the fact that the rules of turn-taking are subject to the conversants' control, i.e. that the rules are designed for being used by each participant to manage their communication with others. By hypothesizing the existence of turn-constructional units, they

1. Goodwin [1981] then goes on to say that on closer inspection things are not as simple as they look. However, the notion argued here is that the principle of turn taking is simple while the behavior emerging from the interaction of the principles of Sacks et al. [1974], especially when observed "in the world," can be quite complicated.

were able to model turn taking with only five—albeit relatively complex—rules. The particulars of the rules are not important here: by far the most important part of their theory is the set of turn-constructional units they propose, which are *sentential*, *clausal*, *phrasal* and *lexical*. These components are used by speakers to construct a turn. For example, recognizing that a particular sentence of type A is being uttered by a speaker, a listener can use her knowledge about sentence type A to predict when it ends, making it possible to take turns with no gaps. However, Sacks et al. fail to specify what kinds of turn-constructional units distinguish one type of utterance—and multimodal act—from another. If we assume that a listener is continuously looking for clues about types, or functions, of utterance segments, a resulting conclusion would be that what is important for extracting these are the features of the utterer's behavior, because, apart from the content of the speech, these are the only clues to the function of the speaker's actions. From a descriptive point of view, turn-constructional units may be valid, but they say nothing about the way people actually recognize these units. What is needed is a mechanism that allows sentential, clausal, phrasal and lexical features to be recognized in real-time and integrated with a discourse participant's actions to allow the pattern of turn taking to emerge. In Chapter 7. we will present a general approach to achieving this.

In what seems to be an incompatible approach, Duncan [1972] proposed the existence of "cues" for turn signalling. It may be argued that Duncan's cues are simply parts of the features that conversants use to identify the turn-constructural units of Sacks et al. In reality, a person uses her perception to make the best or most appropriate decisions at any time regarding her behavior; perception decision is constrained by time, accuracy and the knowledge of the participant. We will come back to this issue in later chapters.

3.1.3 Back-Channel Feedback

Face-to-face interaction quickly breaks down if communication can only happen at or above the turn level [Nespolous & Lecours 1986]—there needs to be a two-way incremental exchange of information. Part of the task for a listener is to make sure that the other party knows that she is paying attention, and indicate that she is at the same state in the conversation. This is done mainly in the back channel [Yngve 1970]. Back channel feedback is in effect information exchange that supports the interaction itself and helps move it along the right path [McNeill 1992, Goodwin 1981]. It includes using paraverbals such as "m-hm," "aha," etc., indicating confusion, expressing feelings (by facial gesture, laughter, etc.), and indicating attentional focus. The absence of such regulatory gestures from a listener may disrupt the discourse [Dahan, as referenced in Nespolous & Lecours 1986].[2] While it may be argued that overlapping talk in the main communication channel is counter-pro-

ductive because it interferes with the flow of a conversation [Sacks 1992b], co-occurring speech in the paraverbal channel does not [Yngve 1970]. The main stream of information (from the speaker) and back channel feedback (from the listener) can therefore be modeled as two separate information channels that can be used simultaneously without interfering with each other. One rule-of-thumb definition of back-channel feedback then is that it is the ongoing (communicative) behavior of a listener that does not change who is in control of the dialogue at the moment.

The above discussion strongly implies that a simple "transmitter-receiver" model will not be sufficient when transferring multimodal interaction to the computer domain. Let us now take a closer look at the role the modes play in multimodal conversation.

3.1.4 Embodied Conversants

Two spatial constraints are of importance to conversation. The first has to do with the location of discussants to each other and the surroundings, referred to here as positional elements, and the second has to do with the conversants' relative orientation, what will be referred to here as directional[3] elements. Surprisingly, research on this topic in psychology is relatively scarce.

Obviously the position of a conversational participant has implications for spatial reference: glances, pointing gestures and direction-giving head nods will be done differently depending on where the speaker and listener are positioned in space. The display of visual cues such as facial gesture is bound to a specific location, i.e. the participants' faces. Multimodal conversants have to be able to find their conversational partners in space—otherwise they would not know where to find the necessary visual information when interpreting each other's utterances or assessing dialogue status. This is important, since a number of turn-taking signals rely on participant location and facial cues [Duncan 1972], and many back-channel feedback cues are given through the face [Goodwin 1981]. Manual gesture are usually given in the area right in front of the gesturer's body [McNeill 1992], and these have to be located in space as well. Gaze is often used to reference this space [Goodwin 1986], and can be indicative of the kind of gesture being made [McNeill 1992, Goodwin 1981].

---

2. Nespolous & Lecours [1986, page 61] say: "... Dahan [see ref., op. cit.] convincingly demonstrated that the absence of regulatory gestures in the behavior of the listener could lead the speaker to interrupt his speech or to produce incoherent discourse."

3. Thanks to Steve Whittaker for suggesting the term "directional."

Orientation has to do with how the participants are turned relative to each other, how various body parts are oriented, and how this changes over the course of the interaction. [Goodwin 1986] For example, turning your head away right after your partner finishes speaking could indicate to him that you think he's done and that you are now preparing a response [Goodwin 1981]. Research has shown that when talking face-to-face, people generally prefer to orient their bodies approximately 90° to each other rather than directly face-to-face [Sommer 1959].

3.1.5 The Multiple Modes of Face-to-Face Interaction

Speech

It has been argued that speech is the main content carrier in face-to-face communication [Sacks 1992a, Sacks 1992b, Ochsman & Chapanis 1974] and may even be the critical medium [McNeill 1992]. Research on language is far more advanced than other aspects of the multimodal interface and is by now a highly mature field compared to other aspects of human communication. Various techniques for parsing natural language have been proposed [cf. Allen 1987]. A clear indication of this is that speech recognizers can now be bought off-the-shelf that are speaker-independent, have a relatively large vocabulary and recognize continuous speech. Researchers have also begun to investigate the link between speech and other aspects of discourse [McNeill 1992, Pierrehumbert & Hirschberg 1990]. For example, McNeill [1992] argues that while on the surface gesture may seem dependent on speech, they often carry different information from the speech they accompany. He proposes that speech and gestures both arise from a common knowledge representation. Pierrehumbert & Hirschberg [1990] have shown how intonation affects the interpretation of speech in context.

The vocal channel is of course also used to give back-channel feedback and other feedback related to process-control. Among the implications of the theory put forth by Sacks et al. [1974] is that turn-taking is a necessary element of any conversational system. They argue that the turn-taking system for speech in fact makes its understanding easier. As a consequence, implementing turn-taking rules and dialogical conventions in multimodal interfaces should make speech communication more robust [Brems et al. 1995], for example by making it easier for the computer to infer where utterances begin and end—still a serious limitation of continuous speech recognition [BBN 1993].

Manual Gesture

In multimodal dialogue, gesture frequently happens along with speech. McNeill [1992] has suggested that gestures and speech are generated

from the same underlying representations in the brain, and others have suggested that the first hominid language was in fact based on gesticulation [Zimmer 1995]. Many classification systems have been used to describe the kinds of gestures people make in discourse [Rimé & Schiaratura 1991, Poyatos 1980], most of them being modifications of Effron's [1941/1972] classification scheme (Figure 3-2). To recap this classification scheme: Symbolic gestures have a direct interpretation in a given culture. An example is the "thumbs up" sign. Deictic gestures are generally referred to as "pointing" gestures. They direct a listener's visual attention to a spatial area or location. To date, symbolic and deictic gestures have been the primary gestures of study at the computer interface (see Table 1). Other kinds of gestures, classified as iconics, beats, pantomimics, metaphorics, tend to carry equally important (and often more complex) information [McNeill 1992, Cassell & McNeill 1991]. Iconic gestures are the kinds of gestures where a body part, often the hands, play the part of another object for the purpose of demonstration. An example would be moving your hand forward, palm down and saying "The car drove like this" meaning that the car moved in some sense the same way your hand does. Pantomimics are gestures where the hand or body of the gesturer are interpreted as real hands. An example is miming the action of hammering or opening a door. Metaphorics are iconic in that they assign meaning to space, but instead of representing concrete objects or events, they present abstract ideas. Beats are rhythmic gestures that accompany speech that have been found to play a large role in the sequencing of turns in dialogue [Duncan 1972], and also to be related to shifts in the dialogue narrative, for example from the main story line to side issues [McNeill 1992]. A last category of gesture is one that should perhaps be classified under "action" instead of being called a "gesture." This is the class of self-adaptors [McNeill 1992, Ekman & Friesen 1969]. Self-adaptors are actions like fixing one's hair, scratching, etc. It has been shown that people attend to such gestures and integrate information conveyed by gesture into their representation of a narrative [Cassell, McNeill & McCullough, forthcoming].

Facial Gesture

Facial gesture has been extensively studied by Ekman & Friesen [1978]. Facial gestures have been found to regulate interaction and they are the primary method, along with intonation, for displaying affect [Ekman 1979]. Pelachaud et al. [1991], following Ekman [1979], classify facial gesture into emblems, emotional emblems, affect display, conversational signals, punctuators, regulators and manipulators. Emblems are movements whose meaning is culturally dependent. An example is nodding for agreement. These gestures correspond to the type of hand gesture that has been referred to by the same term.[4] Emotional emblems convey signals about emotion. The crucial distinguishing feature here is that the gesturer does not feel the emotion at the time of the gesture, but

1. Nondepictive gestures: speech markers (beats)

   A. Stress some elements of the speech for the sake of clairity.
   B. Parallel the introduction of some new element in the discourse.
   C. Chunk the sentence following the steps of the underlying reasoning.
   D. Related: batons, minor qualifiers, beats, paraverbals.

2. Depictive gestures: ideographs.

   A. Sketch in space the logical track followed by the speaker's thinking.
   B. Parallel abstract thinking.

3. Iconographic (iconic) gestures

   A. Present some figural representation of the object evoked in speech.
   B. Subclass: (a) pictographic: represents the shape.
   (b) spatiographic: represents some spatial relation.
   (c) kinetographic: represents some action.
   C. Related: physiographic, motor-primary, illustrative gestures.

4. Pantomimic gestures

   A. Play the role of the referent.

5. Deictic gestures (pointing)

   A. Point toward some visually or symbolically present object.

6. Emblematic gestures (symbolic)

   A. Are devoid of any morphological relation with visual or logical referent.
   B. Have direct translation into words.
   C. Have a precise meaning known by the group, class, or culture.
   D. Usually deliberately used to send a particular message.

FIGURE 3-2. Classification of the kinds of gestures encountered in natural dialogue (after Rimé & Schiaratura [1991]).

merely refers to it via the facial display.[5] Affect display, on the other hand, is the direct expression of emotion. Conversational signals are facial gesture made to punctuate speech, to emphasize it. An example is that raised eyebrows often accompany accented vowels. Punctuators are movements that occur during pauses. Regulators control the speaking turn in a conversation. Manipulators correspond to self-adaptors of hand gestures. An example for the face would be blinking to keep the eyes wet.

Gaze

Most psychological research dealing with gaze has used it as an indirect measure of something else: how long does it take to read a word, what are the mental stages we go through when we try to understand some-

4. Some researchers use the term “symbolic” instead of “emblems”.

5. This classification makes a boolean class out of a continuum, since a facial emotional emblem could be related in any degree to the underlying emotions.

thing three-dimensional, how much resolution do we have in our peripheral vision, etc. The emphasis here is on the role that gaze plays in communication.

Gaze has been shown to be related to a person's attention [Kahneman, 1973], deictic references [Cooper 1974], mental activity [Rayner 1984, Yarbus 1967], and personality, interpersonal attitudes and emotional states [Argyle et al. 1974, Kleinke 1986]. Primarily, gaze is an indicator of a person's attention over time [Kahneman, 1973], and provides therefore crucial information in the conversational setting. People have a strong tendency to look toward objects referred to in conversation [Cooper 1974], which can provide listeners with important deictic information. People will even look where they are listening [Riesberg et al. 1981]. Research has shown that people are extremely good at estimating the direction of gaze of others [Anstis et al. 1969, Gibson & Pick 1963]. The accuracy is dependent on the 3-D aspects of the eyes, the presence of a face around them and the position of the viewer in relation to the eyes [Anstis et al. 1969].

Yarbus [1967] was among the first to show that eye movement patterns vary according to the mental activity of the looker. Subtle differences in gaze pattern were observed to correlate with subtle differences in the task that the looker is engaged in. For example, a picture containing people will be scanned slightly differently depending on whether the onlooker is trying to estimate the people's ages or their wealth. Whether subtle differences like these can be picked up by participants in a conversation is, on the other hand, a question that is difficult to investigate.

Since the eyes are used to gather information, their movements also tell others about this information gathering process. It is therefore not surprising that the eyes also are important in the regulation of turn-taking between dialogue participants [Argyle & Cook 1976]. Argyle and Cook [1976] have shown that the "...gaze patterns of speakers and listeners are closely linked to the words spoken, and are also important in handling the timing and synchronizing of utterances" [p. 98]. They have found gaze to serve three main functions: sending social signals, opening a channel to receive information, and controlling and synchronizing speech. There is a "...very rapid and complex coordination between speech, gaze and other non-verbal signals" [p. 114].

At the initiation of conversation, and during farewells, the amount of gaze between the conversants increases. For the period of the conversation they tend to reach an equilibrium in the amount of mutual gaze. The amount of expected mutual gaze given two speakers' look time can be found by using the following formula [Argyle & Cook 1976, Argyle & Ingham 1972, Strongman & Champness 1968]:

$$EC = EC_1 + EC_2 = \frac{L_T(A) \cdot L_L(B)}{A\text{'s talking}} + \frac{L_T(B) \cdot L_L(A)}{B\text{'s talking}}$$

where EC is expected mutual gaze, L_L represents looking (at other person) while listening, L_T is looking while talking, and A and B are the conversants. In a normal conversation, the average amount of looking at the other person while listening is 75%; the average time spent looking while talking is 41% (given that neither party is trying to avoid or seek visual contact). A and B's look times are determined by the social context (how close people are, who is the other's superior, etc.). Although this formula could hypothetically be used for controlling the gaze behavior of a computational agent or robot by approximating the value of EC in real time during conversation, given the user's gaze input, a more realistic approach would try to model the mechanisms underlying the gaze pattern observed. A number of factors complicate the matter, among them the fact that in addition to being dependent on dialogue state, gaze behavior also varies with the topic of discussion [Cooper 1974]. On top of this lie multiple mental processes influencing the exact observed gaze pattern.

Multimodal Synergism

An important claim of the turn-taking theory put forth by Sacks et al. [1974] is that to get reliable interaction, interactors need to have an understanding of multiple modes. Thus any system that proposes to use turn-taking—as it occurs in human-human interaction—as part of a computer interface will need to incorporate multimodal analysis and interpretation. As we have already mentioned, the flexibility of social interaction stems both from an ability to switch dynamically between representational styles and from combining modes for displaying a single message [cf. Goodwin 1981, Poyatos 1980]. A synergism of multimodal interaction results from the combinatorics of various modes and signals at specific times in the interaction sequence. Any system that tries to introduce flexibility into multimodal human-computer interaction has to take this into consideration. Research on the combinatoric aspect of face-to-face dialogue is still in its infancy [Poyatos 1980] although some guidelines are emerging. Clark and Brennan [1990] present a cost model for combining multiple modes given various constraints in the communication channel. Whittaker and Walker [1991] discuss further the advantages of media types for interface design. The advantages of exploiting the synergistic nature of mode combinations at the computer interface are discussed in Bolt [1987] (see "Face-to-Face: When & Why" on page 26).

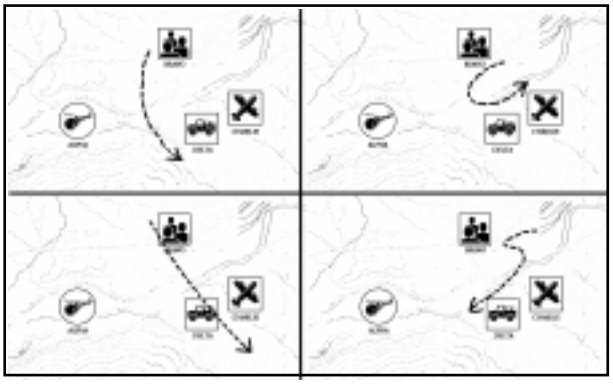


FIGURE 3-5. Example of a multimodal interaction. The user can say "Delete [gesture] these icons" and do a gesture (dotted arrows) near a group of objects. A simulated perceptual grouping algorithm enables the computer to infer which objects the gesture refers to—independently of its precise form [from Thórisson 1994].

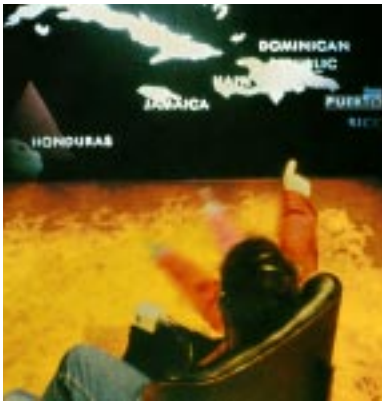


FIGURE 3-3. *Put-That-There* was an early multimodal interface prototype [Bolt 1987].

3.2 Multimodal Computer Interfaces

Having looked at psychological and linguistic research, we now turn to previous computer systems that build on the idea of multimodal, social communication.

In the past, implementations of multimodal computer interfaces have included the use of natural language, either spoken or written, and, to varying degrees, gestural input and eye tracking. A comparison of recent systems is shown in Table 1. One of the first (if not *the* first) system to demonstrate gesture and speech at the computer interface was *Put-That-There*, developed by the Architecture Machine Group at M.I.T. [Bolt 1987, Bolt 1985, 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. Recently there has been an increased effort to combine gestures and language at the interface [Bers 1995a, Thórisson 1995a, 1994, Wexelblatt 1994, Koons et al. 1993, Sparrell & Koons 1994, Sparrell 1993, Neal & Shapiro 1991, Wahlster 1991].

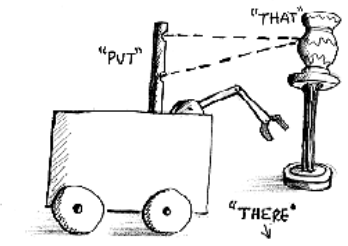


FIGURE 3-4. Cannon [1992] developed a robot that could understand deictic commands by triangulating camera orientation.

CUBRICON [Neal & Shapiro 1991] used typed and spoken sentences as input, along with deictic (pointing) mouse clicks to allow for interaction with a two-dimensional map. A similar system developed at the M.I.T. Media Laboratory [Koons et al. 1993] also uses a two-dimen-

SYSTEM	DESCRIPTION			INPUT				OUTPUT	
Authors	**Goal**	**Metaphor**	**Topic**	**Speech**	**Gesture**	**Gaze**	**Hardware**	**Visual**	**Auditory**
Bolt & Herranz 1992	Manipulation of 3-D graphics	One-way multi-modal	Graphics manipulation	Discrete word recognition	Iconic	Deictic	Gloves, head mic, head eye-tracker	3-D graphic objects	No
Koons et al. 1994	Arranging 2-D icons	Multi-modal dialogue	Firefighting/ 2-D map	Discrete word recognition	Deictic	Deictic	Gloves, head mic, head eye-tracker	2-D map with icons	Synthesized speech
Neal & Shapiro 1991	Information access	Multi-modal dialogue	Military activities	Discrete word recog. Typed NL	Deictic	No	Mouse, keyboard, microphone	2-D map w/ icons, deictic refs., text	Synthesized speech
Sparrell & Koons 1993	Arranging 3-D, graphical objects	One-way multi-modal	Furniture in a virtual room	Continuous recognition	Iconic	No	Datagloves, head mic	3-D graphic objects	No
Starker & Bolt 1990	Interest-responsive storytelling	User as observer, comp. as storyteller	Little Prince's planet	No	No	Deictic, attention	Table-mounted eye-tracker	3-D graphical world	Synthesized speech
Maes et al. 1994	Playful interaction in virtual worlds	Non-verbal interaction	Dogs, creatures and critters	No	Emblems, full body	No	Cameras	3-D graphics	No
Chin 1991	Help for line-command systems	Computer as tutor, user as student	UNIX commands	Typed NL	No	No	Keyboard	Typed NL	No
Jacobs 1990	Object selection	Augmented direct manipulation	Boats on a 2-D map	No	No	Deictic	Keyboard	2-D map with icons	No

TABLE 1-1. Comparison of recent systems that have employed a combination of gaze, gesture and/or speech/NL at the computer interface.

sional map using spoken commands (Figure 3-5), deictic hand gestures, but with the addition of deictic eye movement analysis [Koons & Thórisson 1993]. Starker & Bolt [1990] describe a system that used gaze as an indication of focus of attention and level of interest. Bolt & Herranz [1992] describe a system that allows a user to manipulate graphics with semi-iconic gestures. Koons et al. [1993] demonstrated how gestures can be very efficient for accomplishing many types of spatial manipulations within graphical worlds. Maes et al. [1994] employ a camera to capture the user's behavior and relieve the user from having to "dress up," at the cost of recognizing only symbolic gestures.

At the other end of the spectrum, Cannon [1992] designed a robot that could interpret speech and deictic gestures made with a camera (Figure 3-4). By pointing camera reticles at objects and locations and commanding the robot to "Put that there", the robot used triangulations and planning to execute acts communicated to it in this manner.

Bers [1995b] developed a system that allows the user to combine speech and gesture to direct a bee how to move its wings (Figure 3-6). Rather than mapping the body directly to the wings, the user communicates her intention to the system by saying "Fly like this", showing the wing action with either her arms, fingers or hands. The salient gesture (see below) is mapped onto the bee's body, making it move as prescribed by the user's pantomime. A user can do the gesture before, during or after the speech. The reason for this flexibility is that the system only allows the user to input one kind of gesture, thus bypassing the problem of gesture classification (see page 44).

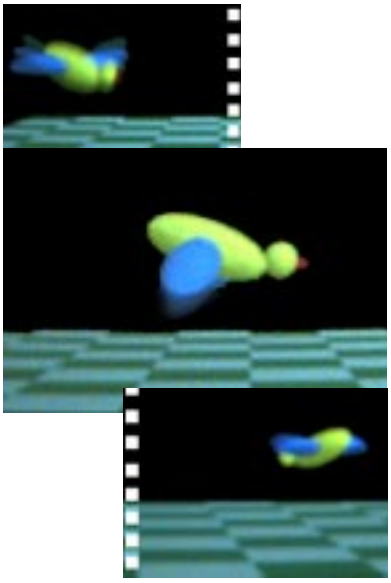


FIGURE 3-6. Bers' [1995b] bee could move its wings according to a pantomimic gesture example provided in a communicative fashion to the system.

Thórisson [1994] began to look at some of the real-time issues of multimodal dialogue by predicting turn constituent boundaries at run-time. This work, which was a precursor to the main contributions of this thesis, is described in detail in Chapter 6., page 81.

3.2.1 Multimodal Analysis and Interpretation

Analysis of Modes

In order to understand, or interpret, a coherent multimodal act, the multiple modes need to be brought together in some way. Such interpretation obviously draws on many resources, including speech recognition, gesture recognition, gaze-following, facial expression analysis, etc. In a free-form interaction, a major problem with gesture is finding which segments are of importance. Sparrell [1993] used a scheme based on a "stop-motion" analysis: whenever there is a significant stop or slow-down in the motion of the user's hand [cf. McNeill 1992, Kendon 1980], the preceding motion segment (called "gestlet" by the author) is

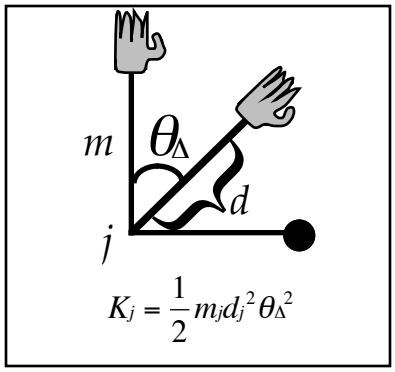


FIGURE 3-7. Kinetic energy of a moving body segment with one degree of freedom is found by looking at its connecting joint: Kj = kinetic energy of joint j, Θ_Δ = difference of joint angle j at t_1 and t_2, m_j = mass of body segment distal to joint j, dj = length of the segment; md = 1 for shoulder, 0.75 for elbow, 0.2 for wrist and 0.25 for fingers. Using a cutoff of 20 units, Bers [1995b] was able to select the meaningful segments of a pantomimic gesture from a stream of body motion data.

grouped and analyzed for lower-level features, such as finger posture and hand position. The interpretation would only happen after the user had finished his utterance. A similar approach was taken by Wexelblatt [1994], adding the ability to refine gesture-interpretation on the fly, as more "evidence" about a motion's trajectory was available. This system was not integrated into a multimodal system, but showed promise.

Bers [1995b] implemented a gesture segmentation scheme based on an orignial idea by the author that utilizes the kinetic energy of body part motion (Figure 3-7). In contrast to Sparrell's approach, this method allows for continuous gesture input. The system computes a "salience" map of body motion (Figure 3-8). Using a cutoff point for the "strength" of a motion, along with time stamping, motions can be selected that relate to the intended speech segment. This method could perhaps be extended to use body motion salience to predict the probability that a gesture is communicative, or to group together symmetric body motions with similar motion strengths.

Recognizing facial gesture has seen some progress in recent years [Essa 1995, Essa et al. 1994, Pentland et al. 1993, Turk & Pentland 1991, Bledsoe 1966]. Essa [1995], employing a camera to provide input to the computer, used optical flow methods to provide an analysis method of facial expression based on Ekman's FACS [1978] model of facial action. This system has not been integrated with other modes or used in an autonomous system.

Automatic intonation analysis has had a very short history and remains for the most part a topic unsolved [Thórisson 1993, Wang & Hirschberg 1992]. This is a problem that needs to be solved in order to create systems that can interact with humans using real-time speech. Speech recognition and natural language understanding have on the other hand been studied for a long time as part of linguistics and computational linguistics [Allen 1987] and will not be treated further here. It suffices to say that current natural language processing systems can have a vocabulary of thousands of words, can be speaker-independent and have a response lag of about 1-3 seconds. The main challenge in these systems remains dealing with brittleness resulting from lack of sensitivity to context and integration of multimodal cues to aid the recognition process.

Multimodal Integration & Interpretation

Koons [Koons et al. 1993] proposed the use of nested frames to gather and combine information from the modes. In his approach the speech is an initiator of gesture analysis: If information is missing from speech (e.g. "Delete that one") the system will search for the missing information in the gestures and/or gaze. Using time stamps, actions in various modes are re-united like pieces in a puzzle, to arrive at a coherent mean-

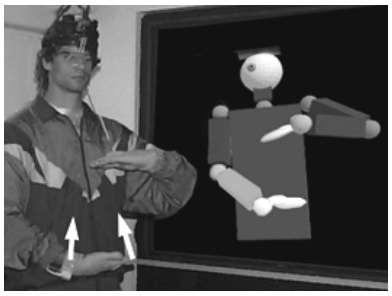


FIGURE 3-8. In this demonstration, salience of the motion of a person's body parts is shown as increased brightness in the corresponding body part of the marionette. Here, using kinetic energy, the gesturer's right-arm gesture (white arrows) is automatically separated from other incidental body motion.

ing. In a functionally similar system, Neal & Shapiro [1991] used a generalized augmented transition network[6] (ATN) that can receive input from a multi-media stream, instead of being limited to linear textual input. This system bypasses the complexities of free-hand gestures by allowing only deixis via a mouse. Others have used a similar method for simplification [Tyler et al. 1991, Wahlster 1991]. However, these put higher emphasis on the complexity of linguistic input allowing, among other things, the use of anaphora.

Compared to machine understanding of natural language, automatic multimodal interpretation is still a relatively undeveloped field. The missing parts include flexibility in interaction, use of cues from one mode to help interpret input from another. This will be discussed more closely in Chapter 5. on the computational characteristics of multimodal dialogue.

3.2.2 Missing Pieces in the Multimodal Metaphor

It may be argued that the main limitation of the multimodal interfaces to date stems from an incompleteness of the metaphor employed. Assuming that face-to-face dialogue is the generic model from which multimodal (and dialogue-based unimodal) interfaces draw [cf. Brennan 1990], one finds that key components are still missing from current implementations. Making this metaphor more explicit will make several things clearer for the user of a multimodal system, as well as for its designer. For example, an invisible, omnipresent agent (as opposed to an embodied one) makes it more difficult for users to pace the interaction and assess its progress at the turn-level [c.f. Clark & Brennan 1990]. Such limitations have been dealt with in various ways; the *Iconic* system [Sparrell & Koons 1994] employs an e-mail style of interaction (construct and send command ..... wait for response) to minimize failures in the interaction sequence. If the interpretation fails, a user has to wait an unknown length of time before re-issuing the command in full. In Wahlster's [1991] system the user selects the desired sub-type of deictic gesture from a menu of icons; the interpretation of the subsequent deictic gesture (a mouse click in a chosen region of the screen) is based on the type of icon selected. Although the interaction in systems such as these can not be called tool-level, it is not fully dialogical either—it seems to occupy a position somewhere between tool-based and communication-based metaphors. As will be argued throughout in this thesis, the critical features of face-to-face communication are not available to a user giving multimodal commands to a computer unless the computer has some command of human multimodal faculties and is explicitly modeled as an interactive agent.

6. See Chapter 6., page 84, for a discussion of the limitation of FSM-style approaches to multimodal interpretation.

To reiterate from Chapter 1, the missing pieces for a fully realized multimodal interface are:

1. Bridging between sensory input and action output,
2. continuous input over multiple modes,
3. integration of this multimodal input (in real-time), and
4. coordination of actions at multiple levels of granularity.

These will be the factors we focus on in the following chapters.