Computational Characteristics of Psychosocial Dialogue Skills

5.

Why is a multimodal interface not just a relapse to the old idea of the teletype where a user would ask the system "questions" or tell it "commands" and the system would reply? Is it really so different? Why, yes. As will be touched on several times throughout this thesis, the difference lies in the interaction itself. Instead of restricting the interface to a vending-machine paradigm, with the call-answer sequence only happening at one level, multimodal interaction calls for a different model where interpretation and action response are intimately tied together, forming a multi-layered interaction space between the user and machine: actions are generated in response to events that happen on various time-scales; these happen in parallel with perceptual and interpretive actions.

```
C:\> cd myfiles\newfiles\cool
Invalid directory
C:\> why?
Bad command or filename
C:\>_
```

FIGURE 5-1. The command line interface has often been incorrectly exemplified as a typical dialogue system.

As explained in the previous chapter, this paradigm contrasts sharply with the computer-as-a-tool metaphor in that the computer is viewed as a dynamic entity, as opposed to a non-acting, dead tool. As pointed out by Laurel [1990], a computer behaves. The actions of computers are sometimes so complex that we cannot not understand them as simply as the light turning on when we flip the switch—we perceive it as if the computer had an agenda of its own. This trend is becoming clearer every year, with ever-increasing complexity at the interface.

Thus, the multimodal metaphor is different both from the old teletype interface and the currently popular object/tool metaphor. When successful, it will feel to the us, computer users, as different as the experiences of hammering in a nail and talking to our children.

We can now begin to take a closer look at the issues behind psychosocial dialogue skills and the unique problems that result from what can be called a holistic approach to multimodal dialogue. We will look at the arguments behind four claims about the process of multimodal interaction:

1. To produce coherent behavior in real-time dialogue, *reactive* and *reflective* behaviors have to co-exist in the same system,
2. analysis of the *contextual function*[1] of speaker actions and control of the process of dialogue are intimately linked through what I refer to as *functional analysis*,
3. the information necessary for *correct and efficient content analysis* is also the necessary information for providing *correct and efficient multimodal feedback behavior*, and
4. tracking of dialogue state should be at the top of the sensory-activities list.

The first relates to the integration of real-time and less-real-time actions, and is dealt with in section 5.2; the second relates to the nature of multimodal processing and is discussed in section 5.3; the third is a derivative of the third and is supported in section 5.3.1. The fourth deals with the priorities of a communicative agent's sensory processes and how this relates to turn-taking, and is found in section 5.4.

In section 5.5 we will look at the important issues of morphological and functional substitutability. At the end of the chapter a layered model of face-to-face dialogue will be presented.

But first we will try to tease out the features of face-to-face dialogue necessary for a computational model.

5.1 Challenges of Real-Time Multimodal Dialogue

Features

From a computational perspective, many features set real-time face-to-face interaction apart from other topics in human-computer interaction and artificial intelligence [Thórisson 1995b]. For the current purposes, these may be identified as:

1. *Incremental interpretation,*
2. *multiple data types,*
3. *seamlessness,*

1. My use of the term "function" is roughly equivalent to its use in speech act theory [Searle 1975, 1969], i.e. as the goal-directed use of communicative acts in context, and is thus close cousin to Austin's [1962] "illocutionary acts", albeit broader. See also Searle's [1971] discussion of function-indicating devices and Silverstein's [1987] treatment of function. The term "contextual" refers to the effect dialogue context, or "state", can have in determining an action's function.

4. *temporal constraints,*
5. *multi-layered input analysis and response generation, and*
6. *functional substitution/morphological substitution.*

1. *Incremental Interpretation.* Multimodal interpretation is not done "batch-style:" There are no points in an interaction where a full multimodal act or a whole sentence is output by one participant before being received by another and interpreted as a whole. Interpretation of multimodal input happens in parallel with multimodal output generation.

2. *Multiple Datatypes.* Multimodal interaction contains many data types, as any quick glance at the human communication modes will show: Gestures [McNeill 1992, Goodwin 1986, Ekman 1979, Ekman & Friesen 1969] provide metric (spatial and spatio-relational information), speech [Allen 1987, Goodwin 1981] provides lexical tokens, semantic and pitch (prosodic) information [Pierrehumbert & Hirschberg 1990], gaze [Kleinke 1986, Argyle et al. 1974, Kahneman 1973], head and body provide directional data related to attention and the dialogue process. These data are both Boolean and continuous over various ranges.

3. *Seamlessness.* When interacting with each other, people generally do not realize that interjecting an iconic gesture into the discourse constitutes a different kind of information than a deictic one, and they don't particularly notice the mechanism by which they take turns speaking. The various data types encountered in face-to-face dialogue have to be combined into a coherent system to allow for seamless multimodal interaction.

4. *Temporal Constraints.* The structure of dialogue requires that participants agree on a common speed of exchange [Goodwin 1981]. If the rhythm of an interaction is violated, it is expected that the violating participant make this clear to others, at the right moment, so that they can adjust to the change. This speed sets an upper limit to the amount of time participants can allocate to thinking about the dialogue's form, content, and to forming responses. (See "Temporal Constraints" below.)

5. *Multi-layered Input Analysis and Output Generation.* In discourse, responses in one mode may overlap another in time, and constitute different information [McNeill 1992, Cassell & McNeill 1990, Goodwin 1981]. The layers can contain anything from very short responses like glances and back channels, to tasks with longer time spans, such as whole utterances and topic continuity generation. In order for purposeful conversation to work, reactive and reflective[2] responses have to coexist to provide for adequate behavior of an agent.

6. Functional substitutability refers to the phenomemon when *identical looking acts can serve different dialogical functions.* Morphological substitutability is the reverse: *Different looking acts can serve the same function.* We will look at this closer in Section 5.5.

Assumptions about the Nature and Quality of Input

When trying to incorporate the above principles into the design of artificial agents, it becomes apparent that certain additional characteristics of the human interpretive processes and quality of "input data" have to be taken into consideration:

1. *Interpretation is fallible.*

Because of inaccuracies in the information delivery of humans, among other things, there will be errors in the interpretation no matter how powerful our interpreter is, whether human or artificial. This problem is worsened in artificial agents by the use of faulty sensors, occlusion when using cameras, background noise masking audio signal, etc.

2. *There are both deficiencies and redundancies in input data.*

It is an inevitable fact that we have to deal with missing information, and, in certain cases, redundancy as a possible solution, both in interpretation and output generation.

3. *Sensory data is collected to allow an agent to produce action or inaction.*

This reflects purpose-directed sensory and cognitive abilities of any situated agent, and prescribes an ego-centered design when producing social behavior in machines.

4. *Behavior is based on data from multiple sources, both internal and external, including dialogue state, body language, etc.*

In multimodal communication action can be taken—and perhaps most often is—based on more than a single piece of information.

5. *behavior is eventually always produced, no matter what data is available.*

Both a listener and a speaker in dialogue are expected to exhibit the necessary behaviors to allow the other to take the necessary steps for clarifying, modifying, and, in general, following the pace of the interaction.

2. This is the issue of how much time one has available for planning a response to a situation. Intuitively, the terms *reactive* and *reflective* refer to fast and slow responses, respectively. There is also a more specific meaning that will become apparent in later chapters. See "A Notation System for Face-to-Face Dialogue Events" on page 108.

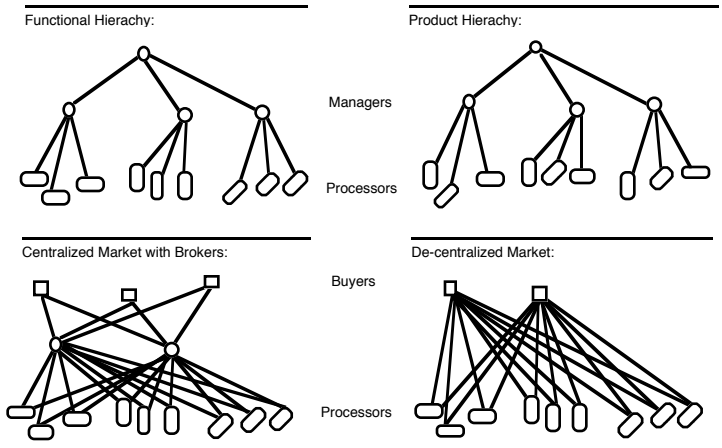
5.2 Temporal Constraints

A useful tool for viewing time-constraints of dialogue is Coordination Theory [Malone & Crowston 1991, Crowston et al 1988, Malone et al. 1988]. Coordination Theory classifies coordination mechanisms broadly into two categories: markets and hierarchies (Figure 5-2). Generally speaking, markets have a relatively high coordination cost and low production cost, whereas hierarchies are the opposite. According to the theory, an object is highly asset specific if it is constrained by extraneous factors, such as place, knowledge, or time. For example, eggs are highly time-specific because they will lose their value if not delivered and consumed before they go bad. Malone et al. [1988] have noted that any highly specific asset is more likely to be handled through a hierarchy. Having seen an example of the time-specificity of dialogue behavior in Figure 1-1 (page 20) it would be natural to choose a hierarchically organized system for its coordination. (We will come back to this issue in Chapter 7.)

As Dodhiawala et al. [1989] have pointed out, real-time performance is not just a matter of speed. They have identified the following four aspects of real-time performance:

- **A.** *Responsiveness:* The system's ability to stay alert to incoming information.
- **B.** *Timeliness:* The system's ability to manage deadlines.
- **C.** *Graceful adaptation:* The system's ability to reset task priorities in light of changes in resources or workload. We should also include under the last part the need to rearrange tasks when problems arise, e.g. the missing of deadlines.

FIGURE 5-2. Four kinds of coordination methods (after Malone et al. [1988]). The mechanisms controlling dialogue behavior are most likely arranged in a product hierarchy.



D. *Speed*

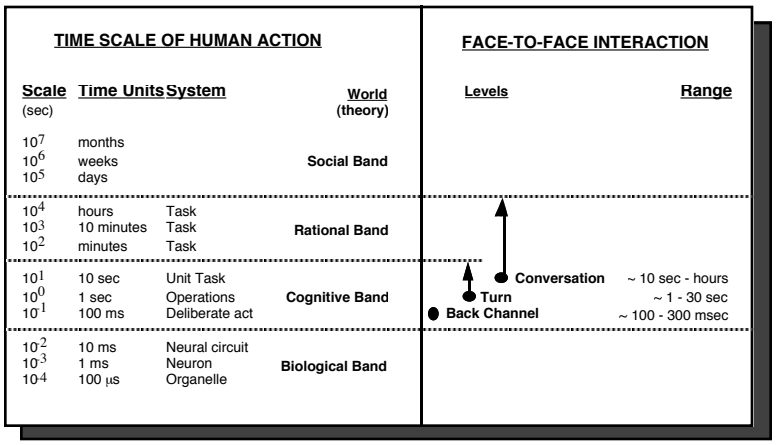
Simply stated, the issue of *speed* may be split into three stages of processing: *speed of analysis* (or perception), *speed of decision*, and *speed of action*. Face-to-face conversation is unique because it contains processes that span as much as 5 orders of magnitude of execution[3] time, from about 100 ms to minutes and hours[4] (Figure 5-3).

As mentioned in Chapter 1., face-to-face discourse [Goodwin 1981] contains rapid responses and more reflective ones interwoven in a complex pattern. This kind of interaction is the basis for the dialogue management system proposed. It calls for an architecture that is responsive to the environment yet is capable of longer-term planning. This is referred to here as "combining reactive and reflective behaviors."

This leads us to claim one:

{1} *To produce coherent behavior in real-time dialogue,* reactive *and* reflective *behaviors have to co-exist in the same system.*

FIGURE 5-3. Comparison between the timing in face-to-face interaction and the time scales of human action as classified by Newell [1990] (from Thórisson [1994]).



3. Notice we are talking about the sense-act cycle, not just motor response.
4. We should say from 0 ms since turn taking is often seen happening with no pauses between speakers. Such phenomena obviously would require some sort of prediction mechanism (if we want above-chance performance) since simple reaction time in humans is typically in the >100ms range and choice reaction time in the >300 ms range [Coren & Ward 1989]. Although it has been shown that prediction mechanism are at work in human dialogue [cf. Sacks et al. 1974], they will not be dealt with here. Suffice it to say that predictive mechanisms could easily fit into the model proposed (Chapter 7.).

This issue is a large one, and one that †mir, presented in Chapter 7. & Chapter 8., provides a solution to.

5.3 *Functional Analysis: A Precursor to Content Interpretation and (sometimes) Feedback Generation*

...if participants are to use each other's bodies as sources of information about their talk they are faced with the task of distinguishing relevant body behavior form that which is not. ...such classification is not simply a hidden cognitive process, but one that has visible consequences for the actions of the party doing that analysis.

—Charles Goodwin (1986, p. 29)

Since any multimodal system works under time-constraints, the natural way to proceed with analysis of the environment is to extract the most important information first. But what constitutes the most important information? How do we look for it? The claim here is that this information is the *function of discoursal actions*, and we look for it using a system of specialized processes that have a relatively high speed/accuracy trade-off.[5]

Initial (basic, elementary) interpretation of a speaker's behavior[6] should not primarily be concerned with what lexical elements can be best mapped onto the user's utterance, or whether the utterance at any point in time is grammatically correct.[7] It should be concerned with distinctions that determine broad strokes of behavior, i.e. extracting the features that make the major distinctions of the dialogue. For example, computing answers to a questions like "is this person addressing *me?"* would be a necessary precursor to start listening. Likewise, answering the question "is the person *pointing*?" would have precede looking in the direction of the pointing arm/hand/finger to find what is being pointed at. These examples constitute analysis of high-level *function*. Computing functions of multimodal actions thus precedes processing the information that is being conveyed.

On the feedback generation side, a listener's behavior of looking in the pointed direction is a sign to the speaker that he knows that her gesture is a deictic one, and that he has correctly extracted the relevant direction from the way her arm/hand/finger are spatially arranged. The gaze behavior resulting from correct functional analysis serves double duty as direct feedback (in this example at least), and constitutes therefore efficient process control.[8]

5. To say that a process ha a high speed/accuracy trade-off simply means that it is more important for that process to provide output in a timely fashion than to be absolutely certain of the accuracy of its output.
6. The claim is made for both computer and human interpretive processes.
7. A similar point is made by Winograd [1988].

Functional analysis—determining the function of a multimodal action—is thus a necessary initial step to both content analysis and correct feedback generation. Let's look at another example, using only the speech mode.

The following exchange may look perfectly fine:

A. *So, aliens ate my Buick.*
B. *I'm so sorry to hear that!*

until we add the accompanying intonation, which goes up at the end of the word "Buick" as indicated with a question mark:

A. *So, aliens ate my Buick?*
B. *I'm so sorry to hear that!*

We find B's response inappropriate and would infer that B thought A was making a remark, not asking a question. If B had "computed" the correct function for A's utterance, (i.e. eliciting information—a question) her response would probably have been different, along the lines of "No, silly!" or "I wouldn't know."

UTTERANCE*	GESTURE**	PROCESS
Speaking	Gesturing	Paying attention
Assertive	Deictic	Addressing me
Directive	Iconic	Giving turn
Commissive	Pantomimic	Taking turn
Declarative	Symbolic	Wanting turn
Expressive	Butterworth♥	
Interrogative	Self-adjustor	
Back channel	Attention-grabber	
Filler♠		

TABLE 5-1. Some main high-level functions of multimodal actions in dialogue. It may be noted that most of a user's utterances directed to an interface agent would probably be directive (commands) and interrogatives (questions).

*See Searle [1975] for a treatment of speech acts. **This applies to both facial and manual gestures [Rimé & Schiaratura 1991, Ekman 1979]. ♠Also referred to as "filled pause;" utterances like "aaaah" and "uuuuuh." ♥The gestural equivalent to filled pauses.

8. It would also be correct and efficient feedback if an agent erroneously concluded that the gesture was iconic and therefore looked at the speakers hand instead, since this would clearly indicate to the speaker the error made. In this case generation of the correct feedback coincides with the actions necessary for further interpretation of the input.

This leads to our second and third claims:

{2} *Analysis of the* contextual function[9] *of speaker actions and control of the process of dialogue are intimately linked through* functional analysis.

{3} *The information necessary for* correct and efficient content analysis *is also often the necessary information for providing* correct and efficient multimodal feedback *behavior*.

The strength of these claims lies in the double support they provide for extracting function before interpreting.

Table 5-1. shows the main functional categories found to date in multimodal dialogue. The issue of functional analysis is neither one of computational power, nor of top-down/bottom-up processing; it is an issue of sequencing. Nothing prevents the use of either top-down or bottom up analysis to extract functional attributes of a speaker's behavior, and adding computational power will certainly speed up the process of analysis. But neither will eliminate the sequential dependency between the two steps of determining an action's function and analyzing its (possible) meaning(s). The second reason why this dependency is important is simple: More assumptions can be made with global information than local—an agent can do a lot more with general information when details are missing than with detailed information when the global perspective is lost.[10] By giving the high-level functions, such as those in Table 5-1., highest priority, the most useful responses can be generated even if other information is missing, resulting in increased robustness.

The functional aspects of multimodal behavior can, and should, be extracted by means of multimodal analysis; any feature, body part, intonational cue or even lexical analysis could assist in the process. A major part of creating multimodal computer agents is finding how to pull out the necessary information.

5.3.1 The Link Between Functional Analysis and Process Control

As we saw in the pointing example, correct and relevant feedback generation often follows automatically from correct functional analysis, but only if we fulfill two conditions. The first is that the behaviors pro-

9. See footnote page 66 for a treatment of "function."

10. This can be seen by a simple example: If I know that I have just been asked a question, but missed some of the words, I can exclude all forms of utterances except questions from consideration and ask the speaker to repeat, increasing the probability of correct interpretation significantly. This does not work in the other direction.

duced by the system be guaranteed execution within a given time limit, as determined by the pace of the dialogue. Because dialogue state is constantly shifting, we need a mechanism that ensures that behaviors be executed at the time they are relevant—not before and not after.

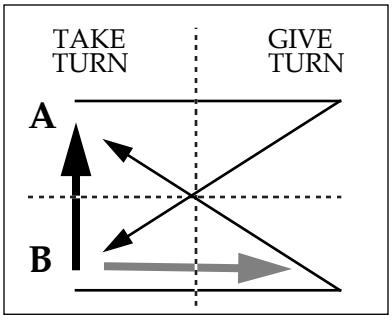


FIGURE 5-4. The problem of efficient turn taking includes detecting the correct transition points. In this figure, A and B are participants in a dialogue. Thin arrows demonstrate smooth turns; solid bold arrow constitutes an interruption (of B by A) with the possibility of overlapping speech, gray bold arrow shows a failure of the listener (A) to take the turn when it is given (by B), possibly with an unwanted silence.

The second condition we need to fulfill is that we model the agent in our own image, i.e. with a head, face, gaze, arms, hands, and a body—organs that have to do with communication. This is because in face-to-face interaction, sensory organs, bodily constraints, attentional and other mental limitations are linked together in a way that is intimately integrated and intertwined with the dialogue process. This provides dialogue with an intricate feedback mechanism, the absence of which has been shown to disrupt discourse [Nespoulous & Lecours 1986].[11] In other words, if any parts of this mechanism are broken or missing, dialogue may break down.[12] It may be added that providing an agent with misleading actions or visual features may of course lead to the same results.

5.4 Turn Taking

As we established in Chapter 3., a key element of dialogue is turn taking. Sacks et al. [1974] maintain that the purpose of the turn taking system is to minimize overlapping speech and pauses in interaction. When we refer to the *seamlessness* of dialogue, we are essentially referring to a collection of mechanism grouped under this hat. Figure 5-4 shows the possible outcomes of turn taking with two participants. The difficulty of correct turn taking by machine lies first and foremost at the perceptual/knowledge level, because a participant has to infer what constitutes a valid turn-giving signal for each role. Generating that signal for the speaker is, on the other hand, simple. So computational turn-taking modelling is first and foremost a perceptual problem.

As numerous researchers have shown [Walker & Whittaker 1990, Goodwin 1986, Sacks et al. 1974], turn taking defines the two main roles of conversants: listener and speaker. Each role calls for its own repertoire of behaviors and perceptual tasks. My proposal is to define two very different classes of behaviors, both of which include percep-

11. Nespoulous & 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."

12. It is also possible that some violations can be fixed with clever engineering of the agent behavior, its visual appearence or its environment. This topic will be revisited in Chapter 11.

tual, decision and motor tasks, that participants in a dialogue have to switch between. Thus, for the period that person A takes the role of listener, one can expect him to be engaged in a set of mental activities—mental activities that are different from those he is engaged in when in the role of speaker. To take an example, Goodwin & Goodwin [1986] discuss the activity of searching for a word and how this can be a cooperative activity. A speaker may indicate to her listener, using gaze and body language, that she is looking for a word. The listener will offer to assist in the search by interjecting plausible words. Although the process is cooperative, it is the speaker who has the turn, and thereby the power to accept or reject the listener's suggestions. In each role it is not only the behavioral repertoire that is different but also the demands on the participant's perceptual and decision-making systems. The roles can be thought of almost as roles in a play; they are part of the same plot but the rules for each character are very different. According to this proposal, a speaker's perceptual system is preoccupied with monitoring the progress one is making in the narrative production of output, what little is left of attentional capacity is spent distinguishing between acts of the listener that are insignificant to the dialogue (such as the listener scratching himself) or constitute communicative actions, such as a wish to interrupt. The listener's role revolves around interpreting what the speaker is saying and making sure she knows that he is following her, as well as interrupting when problems arise.

This emphasis on the listener-speaker distinction has the important effect of putting the tracking of dialogue state at the top of the sensory-activities list. It is a process that happens at the decisecond level of granularity (see Figure 1-1 on page 20) and as such has a relatively high speed/accuracy trade-off—i.e. it is highly temporally constrained.

This leads us to claim four:

{4} *Tracking* dialogue state *is at the top of the sensory-activities list.*

But what should these sensory activities be?

5.4.1 A Situated Model of Turn Taking

The model of turn taking advanced by Sacks et al. [1974] is very broad and can be considered to be about as good as a descriptive model of turn taking can get by just using data from human observation and video tape analysis. To design a system that can actually generate turn taking behavior and exhibit the rules described in their model, one needs to make several decisions about the nature of the underlying perceptual mechanisms. Here, two hypotheses are put forth for making the creation of such a system possible.

Earlier we claimed that functional analysis of multimodal actions is necessary for providing correct multimodal feedback (page 71). The need for reactive responses puts definite time constraints on this analysis that have to be met for the system to work. This leads us to hypothesis 1:

1. Reactive behaviors[13] are based on opportunistic processes: *High-speed functional analysis in multimodal dialogue draws on cues from any number of number modes, as long as they are informative*.

We also need to specify how the information from various modes is combined. The second hypothesis is this:

2. *Features extracted from a particular multimodal speaker action are logically combined (in the mathematical sense of the word) by the listener to arrive at a plausible dialogue function for that action* [cf. Duncan 1972].

To relate this back to the issue of the speed/accuracy trade-off in perception and action, according to these hypotheses, an increased number of features and modes included in a single analysis will strengthen the *accuracy* of that analysis, but do not affect its *speed*. Thus, the reliability of the turn-taking process should increase with an increased number of cues, but the speed of analysis will not change. However, and here is the rub, increased reliability may affect the speed at which the extracted functions will be *acted on*. Thus, upon interpreting the multimodal act "He went [deictic manual gesture & gaze] that way," a listener may look sooner in the relevant direction if the manual pointing gesture is present, than if she only has gaze as an indication of direction, since a manual deictic gesture is a more reliable indicator of direction than gaze alone.

No claims are made here whether this model is "true"—that remains to be answered by experimentation. It enables us, however, to start building a system that allows such experimentation to take place.

5.5 Morphological and Functional Substitutability

- Morphological[14] substitutability[15]: *Different looking acts can serve the same function.*
- Functional substitutability: *Identical looking acts can serve different functions.*

13. The terms reactive and reflective are dealt with in Section 5.2, page 69.
14. The term "morphological" is used here as relating to "form".
15. My thanks to Steve Whittaker for the term "substitutability".

One of the problems of multimodal interaction is the variability in the way people communicate the same meaning. To take an example, a pointing gesture can be made with a head nod, a wave, a pointing finger, etc. Is the list infinite for any given function? No, clearly, if it was, people couldn't understand each other. The assumption here is that the morphology—or certain features of the morphology—of a multimodal action is mapped to its function by social convention. We can call this the *morphological-functional* link. Thus, for any given society, there are approved ways of communicating certain information. We can choose any of the above ways for pointing out an object in our surrounding. This phenomenon is referred to here as morphemic substitutability. If we want to be as clear as possible when pointing out an object in the environment, the best way to this in English speaking countries is with an extended arm and extended index finger, with the index finger pointing approximately in the desired direction. Clearly, this is not only a matter of intelligent use of the human figure to convey information, but also a matter of establishing a morphological-functional link. The more sloppy we are in extending the arm and finger, the more noise we introduce into the communicative act. This, then, suggests a second property of morphological-functional mapping: A graded index of flexibility, where certain morphologies are more strictly mapped to function than others. Mapping from morphology to function is strictest for words and symbolic gestures, and most flexible for sentences, speech acts and iconic gestures.

The corollary to morphemic substitutability is functional substitutability, where identical looking (or sounding) acts can serve different functions. An example is scratching your face while listening versus scratching your face when talking about an itch you had yesterday. The functional extraction in these cases has then to proceed by other indicators than morphology, the primary ones being content, dialogue state and the attentional state of participants [Grosz & Sidner 1986].

5.6 *Multimodal Dialogue as Layered Feedback Loops*

The model put forth here of multimodal interaction can be characterized as a layered feedback-loop[16] model, and is intended to be *de*scriptive—

16. "Feedback" in this context refers to the reciprocal nature of any speaker-hearer relationship, where a participant's [P1] multimodal action [P1-1] is met by the other's [P2] re-action [P2-1]. This loop can be more than one level deep; a common format is the sequence [P1-1→P2-1→P1-2]. See e.g. Clark [1992].

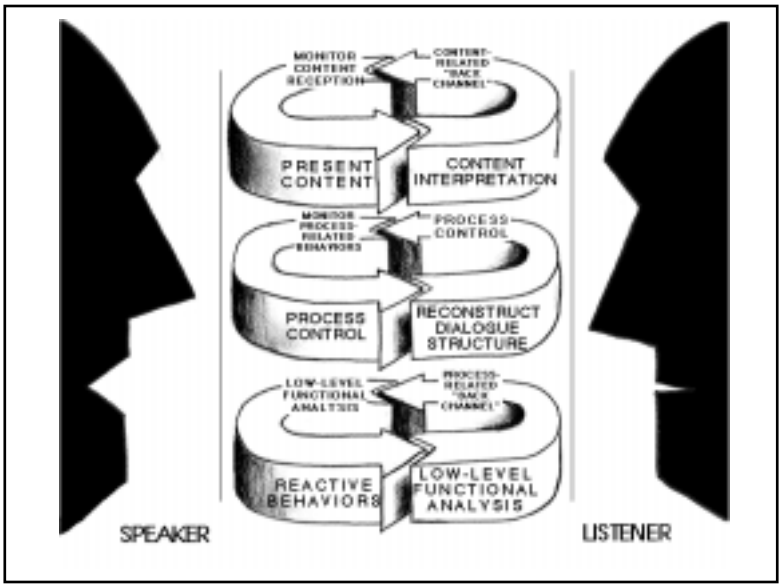


FIGURE 5-5. The proposed three-layered model of multimodal dialogue.

it is based on research from the psychological and linguistic literature—as well as *pres*criptive—it specifies how a conversant can be constructed (Figure 5-5). The three layers in the model are based on the time-scale of actions found in face-to-face dialogue (Figure 5-3). At each level various sensory and action processes are running, thus belonging to the category of functional hierarchy in Coordination Theory [Malone & Crowston 1991, Crowston et al 1988, Malone et al. 1988]. The set of sensory and action processes at work in each layer at any point in time is mostly determined by the role of the participant at that point in time: speaker or listener.

The lowest level is concerned with behaviors that generally have recognize-act cycles shorter than 1 second. This is the Reactive Layer. The middle layer concerns behaviors that are usually slower than 1 second. This is the Process Control Layer. Together these two layers define the mechanisms of dialogue management, or psychosocial dialogue skills. Direct references to the process of dialogue—e.g. utterances like "I'm trying to remember..." and "Let's see..."—belong in the Process Control layer and are generated in response to the *status of processes* in the other layers. Highly reactive actions, like looking away when you believe it's your turn to speak [Goodwin 1981] or gazing at objects mentioned to you by the speaker [Kahneman 1973], belong in the lowest layer. The third part of this model is the Content layer, where the content or "topic" of the conversation is processed. This layer deserves its own discussion,

and will be dealt with more in Chapter 7. and Chapter 8. The layers will all be more closely examined in these sections as well.

5.7 Summary

In this chapter we laid the foundation of a computational framework for face-to-face interaction. We identified the issues of real-time interaction that have to be solved for a satisfactory computer model as being [1] incremental interpretation, [2] multiple data types, [3] seamlessness, [4] temporal constraints, and [5] multi-layered input analysis and response generation. A proposal was made for a well-defined distinction between processes responsible for the behaviors of listeners and speakers. It was maintained that analysis of the function of multimodal acts has to happen before content can be successfully extracted from any such act. we presented the concepts of morphemic and functional substitutability—based on the observation that different looking acts can serve the same dialogical function, and that identical looking acts can serve different functions. The morphological-functional link is the proposal that morphology of an act is mapped to function by *social convention*.

Lastly, a proposal was also made for multimodal dialogue as semi-independent layered feedback loops, with each layer being responsible for separate parts of sensation and perception of an agent.

We will now see where this groundwork leads to: The next chapter will look at J. Jr.—a pilot study in face-to-face reactivity, and then, in Chapter 7., present †mir, a model for the generation of real-time multimodal dialogue behavior.