Int'l J. of Human-Computer Studies, to appear March 31, 2003

short running title: Responsiveness in Spoken Dialog

A Study in Responsiveness in Spoken Dialog

Nigel Ward[1] and Wataru Tsukahara[2]

Mech-Info Engineering, University of Tokyo
7-3-1 Hongo, Bunkyo-ku, Tokyo 113-8656 Japan

Abstract: The future of human-computer interfaces may include systems which are human-like in abilities and behavior. One particularly interesting aspect of human-to-human communication is the ability of some conversation partners to sensitively pick up on the nuances of the other's utterances, as they shift from moment to moment, and to use this information to subtly adjust responses to express interest, supportiveness, sympathy, and the like. This paper reports a model of this ability in the context of a spoken dialog system for a tutoring-like interaction. The system used information about the user's internal state — such as feelings of confidence, confusion, pleasure, and dependency — as inferred from the prosody of his utterances and the context, and used this information to select the most appropriate acknowledgement form at each moment. Although straightforward rating reveals no significant preference for a system with this ability, a clear preference was found when users rated the system after listening to a recording of their interaction with it. This suggests that human-like, real-time sensitivity can be of value in interfaces. The paper further discusses ways to discover and quantify such rules of social interaction, using corpus-based-analysis, developer intuitions, and feedback from naive judges; and further suggests that the technique of 'evaluation after re-listening' is useful for evaluating spoken dialog systems which operate at near-human levels of performance.

1 INTRODUCTION

1.1 Real-Time Social Interaction

There is something nearly magical about human-to-human interaction. When a conversation goes well it can be very pleasant indeed. You may achieve a sense of being 'in synch' with the other person, of having 'connected', or of being 'on the same wavelength'. These benefits of human-to-human dialog are, to a large extent, obtained orthogonally to the "official business" (Clark, 1996) of the dialog. Even if there is no real content, as in talk about the weather, or

[1]currently with the Department of Computer Science at the University of Texas at El Paso. El Paso, Texas 79968-0518 USA. nigel@cs.utep.edu

[2]currently with the National Institute for Japanese Language Research, 3-9-14 Nishigaoka, Kita-ku, Tokyo, Japan

I want to express my thoughts *(by taking a turn soon)*	I am in the midst of expressing a thought *(so please listen and don't interrupt)*
I'm uncomfortable *(with this topic)*	I'm pleased *(that you appreciate the irony in my words)*
I'm amused *(by your story)*	I'm not committed to any opinion or plan *(so you're welcome to keep making your proposal)*
I'm frustrated *(that I'm expressing myself poorly)*	I'm bored *(so let's talk about something else)*
I'm happy *(to just keep listening)*	I'm concerned *(that I'm not expressing myself well enough)*
I'm missing something *(so you need to be more explicit)*	I'm really interested *(in your opinion on this)*
I need a moment *(to digest that statement)*	I know just how you feel about that *(and sympathize)*
I know what I'm talking about *(so please believe me)*	I'm knew that already *(so we can go on to talk about something else)*
I'm surprised *(at that new information)*	I'm getting restless *(so let's close out this conversation)*
Whatever, I don't care *(so don't expect me to pay attention)*	I'm feeling a twinge of irritation *(at the tone of your last remark)*
I'm expected you to say that *(so I'm with you, go on)*	

Table 1: Examples of Feelings and Attitudes that Occur in Dialog, as suggested by studies of prosody, back-channel lexical items, disfluency markers, and gestures, as they occur in tutorial-like dialogs, casual conversations and narrations (Bavelas *et al.* 1995, Ward and Kuroda 1999, Ward 2000)

divisive content, such as a difference of opinion, the same feelings of satisfaction can arise. Thus the *process* of dialog itself, not just the task served, can be valuable.

While there are various factors contributing to the pleasure found in dialog (Bickmore and Cassell, 2001), one component, we believe, is the successful exchange of information about the participants' states, in real time as they change moment-by-moment during the dialog. That is, it can be satisfying when a conversational partner accurately tracks the attitudes and feelings which subtly color most of our utterances. Examples of these attitudes and feelings are seen in Table 1. Of course, not all listeners bother to, or are able to, track these attitudes and feelings, but when someone does, and shows this with their face or voice, it can be very satisfying.

Dialog is rich in non-verbal signals, and many of these, we believe, relate to these attitudes and feelings. Insofar as dialog participants generally send and receive these signals without conscious attention, it possible to explain why dialog can seem magically pleasant.

This paper addresses the question of whether it is worthwhile for spoken dialog systems to be human-like in this way: that is, whether users appreciate a system which picks up these subtle signals in order respond appropriately to aspects of the user's internal state.

1.2 Prospects for Spoken Dialog Systems

Speech technology has recently reached the point where there exist systems which allow motivated users to accomplish useful tasks.

However, interacting with these systems is still a far cry from interacting with a human[3]. In order to build dialog systems for a broader range of applications, it is necessary to consider how to make them effective for users who are less motivated. That is, in many situations, it will be important for interaction with the system to itself be pleasant, otherwise the customer may hang up before making the purchase, or the student may exit the application before he gets to the point of learning anything. Putting it more positively, systems in many domains will need to be able to motivate, charm, persuade, and amuse users.

For this, accurate speech recognition and efficient dialog management are necessary but not sufficient. To obtain the necessary quality-of-interaction will require something more, what we have been calling 'responsiveness' (Ward, 1997) — the ability to speak at precisely appropriate times with precisely appropriate utterances.

Although there is a substantial body of research on the usability aspects of speech interfaces, the problems of achieving responsiveness have received little attention. This is largely because today the problem of speech recognition itself is so hard as to overshadow all other considerations in spoken dialog interface design. Design is currently limited by the need to constrain the interaction in various ways so as to allow the recognizer to perform accurately. As a result dialog system design today is mostly a matter of working within these constraints to avoid gross infelicities, of mis-recognition, mis-prompting and so on (Walker et al., 1998), and of using various counter-strategies for achieving relatively natural interaction (or at least tolerable task-completion rates) despite these constraints (Yankelovich et al., 1995; Mane et al., 1996; Oviatt, 1996).

1.3 Overview of the Paper

Thus, if we want to build dialog systems that people will find pleasant to interact with, or at least less tiresome, we probably should attempt to model responsiveness in real-time social interaction. In particular, we need to discover the signals used and to exploit them in systems which can detect and model the user's internal state, and respond appropriately. If we can do this, it should be possible to engineer systems that seem responsive, easy to talk to, and perhaps even sympathetic, supportive, or charming. This paper describes a investigation of this possibility. Section 2 presents the domain chosen, mock-tutorial dialogs, and observations of how people interact. Section 3 explains how we analyzed what makes some tutors successful in this domain. Section 4 presents the rules we discovered for chosing acknowledgements suitable for the user's internal state. Section 5 explains the experimental method. Section 6 presents the result: that users preferred the more responsive system when ratings were solicited appropriately. Section 7 discusses the probable generality of the finding, and relates it work on exploiting emotion and

[3]Here as elsewhere we will treat human performance as the ideal to which dialog system designers should aspire, although in reality there are many cases, such as dealings with California Income Tax Telephone Assistance, where human performance levels are below the level required to make conversations feel magical or even pleasant.

social conventions in interfaces. Section 8 discusses the prospects for developing such systems cost-effectively and some open research questions. Section 9 summarizes.

2 TUTOR SUPPORT IN MEMORY GAMES

To study these phenomena we chose one of the simplest dialog forms imaginable: practice memory quizzes. If you need to memorize chemical symbols, multiplication tables, geographical names, or the like, one technique is to have someone quiz you verbally, as a way to test your knowledge and to motivate you to study further. Of course, this only works if the tutor is helpful, sympathetic, encouraging, and fun: otherwise it can turn into a dull chore. Thus it seems that tutoring is valuable not just due to the efficient conveying of facts, but that "there is something about interactive discourse [itself] that is responsible for learning gains" (Graesser et al., 1999).

Our initial interest in this domain was as a task in which it would be possible to produce a spoken dialog system able to "keep up with the user". Most dialog systems today have a fairly rigid turn-taking structure. The system speaks, there is a pause, the user speaks, there is a pause, and so on. Human-human interaction is generally not like this[4]. For several years we have been working towards the realization of a more normal, more pleasant form of interaction. Our first step was the development of a system able to respond more swiftly, overlapping the user's utterances when appropriate (Ward and Tsukahara, 1999). The system, although only capable of controlling the timing (of back-channels), seemed to provide a qualitatively different sort of interaction: much more intense and involved. As the next step we chose memory quizzes, because of the clear need for swift interaction, and because it seemed to be a domain where the speech recognition problem would be tractable, allowing the possibility of a convincing demonstration of the value of responsiveness. Although our goals shifted during the course of the project, this choice of domain turned out to be a good one.

The specific task we chose was a simple one, where one person, playing the role of tutor, prompts the other to "try to name the stations on the Yamate Loop Line, in order". The other person, playing the role of student, does his or her best to recall them. After each guess, the tutor checks the answer against the map, tells the student whether he or she was right, and if not, gives hints. There are 29 stations on the Yamate Loop Line; the average Tokyo-ite knows many but requires hints for the rest. Comparable tasks include naming the 13 original United States or the 15 countries of the European Union.

Viewed in terms of information content, such dialogs are trivial: the student produces guesses, and the tutor indicates whether the answer was correct or incorrect, and if incorrect provides a hint. But dialogs for this task are actually quite rich and varied. More often than not, participants seem to find these little dialogs enjoyable.

As suggested in the introduction, this seems be be, in part, because tutors vary responses depending on the feelings of the student interlocutors. A good example is seen in Figures 1 and

[4]Some recent systems allow the user to "barge-in", interrupting the machine and making it shut up, which does allow more efficient exchanges, but is nevertheless not a form of turn-taking common in polite interactions.

S: Shibuya eetoo Gotanda a Ebisu eeto Ebisu, Gotanda?	
T: hai buu hai buu	
S: Ebisu? Ebisu no tsugi? nani ga aru?	
T: Ebisu no tsugi ha? hora Mejiro ja nakute	
S: a Meguro ka	
T: haihai	

S: Shibuya let's see Gotanda oh, Ebisu let's see, Ebisu, Gotanda?	
T: yes bzzzt yes bzzzt	
S: Ebisu? what's after Ebisu? what's there?	
T: after Ebisu is? come on not MeJIro, but	
S: oh Meguro, maybe	
T: right	

Figure 1: Sample Human/Human Dialog from the Corpus. Here **S** is trying to remember the names of some train stations, and **T** is trying to help him. Japanese original above; English translation below

S: ··· HAMAMATSUCHO no tsugiwa	NIHONBASHI!
T: saishoni resshaga hashittatoko	Bu-!
S: aha chigau? e-to SHINBASHI!	
T: soso!	

S: ··· after HAMAMATSUCHO is	NIHONBASHI!
T: the first train station	No!
S: (laugh) no?, ohh SHINBASHI!	
T: That's it!	

Figure 2: Another Fragment of this Dialog.

2, transcriptions of a human to human dialog from our corpus. In this example, **T** seems like a helpful buddy. He listens well; he does not interrupt **S** when things are going smoothly, but he gave a helpful hint when **S** is in confusion. And when **S** blurts out an answer after some struggle, **T**'s response is lively; he seems to be sharing **S**'s moment of pleasure.

While this task is a simple one, these phenomena seen in these little dialogs are also present in much more complex conversations. For example, customer-salesman interactions also exhibit times when one participant is supporting and giving feedback, for example when the customer is prompting and encouraging the salesman to recall some necessary technical detail, or when the salesman is helping the customer to recall or determine what the key requirements are.

We chose to study these phenomena by building a a system to take the tutor's rule in such interactions.

3 ANALYZING REAL-TIME INTERACTION

There have been few, if any, previous attempts to analyze dialog at the level of detail required to build a system faithful to real-time human behavior, and so there is no standard way to go about doing so. This section describes the methods we used, without pretending that this is the only or best way.

3.1 Focusing on Acknowledgement Choice

Our first step was to record 41 dialogs of people doing the Yamate-sen game, recruiting co-workers and friends in pairs to play the role of student and tutor, for a total of 146 minutes of data.

We then listened to the dialogs, repeatedly, to determine what aspects were worth modeling and implementing. We chose to focus on variation in acknowledgements, since this was the most common way in which the tutors seemed to make the dialogs fun. Other aspects, such as choice of hints, were less interesting and general. The aspect of acknowledgements which we chose to focus on was word choice, since variation in acknowledgement timing and prosody, although significant, seemed less expressive and less varied, and also seemed harder to analyze.

Thus our task was to model the rule by which the tutor chose how to respond to correct station names. The choice is analogous to the choice in English between *yes*, *that's right*, *right*, *yeah*, *okay*, *uh-huh*, *mm-hm*, echoing back the correct station name, and remaining silent. Unfortunately, there is no simple correspondence between the acknowledgements in the two languages, so below we will just discuss the Japanese choices.

In focusing on acknowledgement choice, our work is similar to that of Graesser *et al.* (1999, see also (Rajan et al., 2001)), who modeled the roughly analogous English items in varying intonations. In this system choice among acknowledgements is determined by the correctness of the student's action, specifically the "the quality of the set of assertions within a conversational turn". Thus this system analyzed only the content of the user's input, not the way the user felt when he or she produced it.

We therefore looked for signals from the users which affected acknowledgement choice[5]. We chose to analyze only audio, since spoken interaction was all we intended to implement. Of course face-to-face interaction is more interesting, but responsiveness has value even over the telephone, and limiting the modality made the problem tractable.

[5]We could equally well have proceeded in the other direction, starting with the non-verbal signals in the user's voice. That is, since variation in the way users say things is generally not just random, we could have attempted to classify the sorts of things they were trying to convey (not necessarily consciously), and then looked for ways in which the tutor did or should have acknowledged or reacted to these. However the strategy we adopted, of starting from variations in the tutor's acknowledgements, gave us a clearer goal.

3.2 The State of the Art

In an attempt to find the signals and rules governing acknowledgement choice we first surveyed previous work. Although there was some very useful information, as described in the next section, in general this aspect of acknowledgements has not been well studied.

While dictionaries do list acknowledgements, they do not discuss the differences in meaning among these items; although this is not surprising, considering that dictionaries primarily catalog the written language, not the spoken language.

While there has been some research on Japanese acknowledgements (Angles et al. 2000), the findings were not specific enough to be helpful. In general, it seems that linguists and conversation analysts have a tendency to study rich dialogs, in which too many confounding factors are present to allow the positive identification of any one.

The work of Shinozaki and Abe (1998) was almost what we needed, in that they looked at the connotations of various possible ways a system could respond to user statements ("I like cheese." "Oh".), and in that the work was on Japanese. However their study did not address acknowledgements as they occur in an ongoing conversation, and so was not directly helpful.

There is a substantial body of work on the correlates of emotion, attitude, dominance, affiliation, etc. in speech signals, surveyed by Murray and Arnott (1993) and Cowie et al. (2001) . However most of this work has used staged emotions, and focused on fairly static emotions, lasting over a few minutes or an entire discourse. The only clear exception is the work of Cowie et al. (1999), which explicitly addresses the problem of identifying emotions which vary second-by-second. Unfortunately this project has so far focused on developing frameworks and tools, and so was not directly helpful.

3.3 Focusing on One Person

We thus had to discover the rules ourselves. Since the corpus contained plenty of data, more than a thousand acknowledgements in all, we planned to use machine learning algorithms to extract the rules automatically. It soon became clear, however, that the behavior patterns of the various people in the tutor role were very different. For example, some tutors produced nothing but *hai* (the most formal acknowledgement, corresponding roughly to English *yes*), others showed a little variation, while others employed a rich repertoire.

Thus we concluded that any 'average' interaction strategy would be intolerably bland, at best. Moreover, although most people in the tutor role performed adequately, there was only one 'great tutor', one who was always responsive and moreover seemed to have enjoyed the dialogs and to make things fun for his 'students'. We decided to base the system on this individual's behavior patterns — to give the system some basic elements of his interaction style. In general this is probably a good strategy: since there exist all sorts of human communicators — personable, effective salesmen along with annoying drones, and skilled private tutors along with nagging nuisances — it makes sense to model the best.

We therefore solicited a further 5 dialogs with this individual in the tutor role (30 minutes

total) and analyzed these. His conversation partners were diverse: 4 males and 2 females, 4 of lower social status, 1 equal and 1 higher — however sex and age did not seem to be influencing his acknowledgement behavior much, so we treated the data as a single set for analysis.

In retrospect it might have been better to have gathered even more data. However at the time we wanted to study "natural" conversations, collected without informing the participants (until later) of the true purpose of data collection. For this reason we did not record more conversations, for fear that our tutor might catch on to the purpose of our data collection and change his behavior, or get bored and less effective.

3.4 Discovering the Rules

Initially we looked for correlations between properties of the input (the context and prosody of the student's correct guess) and the output (the system's acknowledgement). In particular, we computed correlations between acknowledgment choice and 88 possible predictive features, such as speech rate, pitch slope in the last 200 milliseconds, and number of wrong guesses over the past three stations. Selecting the strongest correlations and simplest features, we hand-coded an initial set of decision rules, as seen in Table 2.

response	condition
hai	previous acknowledgement was *hai* and no previous incorrect guess
<echo>	long delay before correct answer
un	station is an obscure one, a hint was given, or guess preceded by a filler
hai	default

Table 2: Preliminary Rules for Response Choice

We then ran these rules, with the input being the students' sides of the dialogs, to generate predicted acknowledgements. We repeatedly refined the rules until their output was, in most cases, the same as the actual tutor's acknowledgement in the corpus.

We then began a second process of refinement. We synthesized conversations embodying these rules, by audio cut-and-paste, and played them to 3 friends not familiar with our research. These "judges" were able to point out cases where an acknowledgement seemed inappropriate for the flow of the conversation, or unnatural, or cold, and so on. Based on these comments we revised the rule set again (Tsukahara, 1998).

When the judges' comments suggested changes that were counter to what we saw in the corpus, we generally favored the judges comments. As a result, the final version performed worse, in terms of corpus-matching accuracy, than the preliminary set of rules, but it sounded better to the judges, and to us. One major quantitative difference was that the final version produced more variation, using the default *hai* less often.

4 RULES FOR RESPONSIVENESS

This section describes our 8 basic rules for chosing acknowledgements, presented roughly in order of discovery. It also explains the reasoning behind their adoption.

4.1 Rule a: Avoid Unnecessary Responses

When someone is "on a roll", swiftly reciting station names in sequence, acknowledgement responses are unnecessary. This was pointed out to us by one of the judges listening to cut-and-paste dialogs. Such cases are also evident in the corpus. In the example below, the tutor waited until the student paused, then acknowledged all the stations names together. Acknowledging each station individually sounds odd.

S: OKACHIMACHI-AKIHABARA-KANDA-TOKYO	
T:	OKACHIMACHI-AKIHABARA-KANDA-TOKYO

In these cases we decided to omit the individual acknowledgements and just produce a single *hai* (*yes*) at the end. This was done by simply adding a rule for the system to suppress output if the student had already started to utter the next answer before the time when the system was to respond.

4.2 Rule b: Respond Cheerfully to Lively Answers

A judge made the comment that the tutor should give a lively response when a student's answer was lively. The notion of "lively answer" seemed to refer to utterances with high average pitch or power. After adjusting the parameters to give a good agreement with subjective ratings, we defined "liveliness" to be:

$$(\text{liveliness}) \equiv \overline{f}_{0norm} + 1.5\overline{E}_{norm} \quad (1)$$

where $\overline{f}_{0norm}$ is the average pitch in the answer normalized by the median pitch for this speaker, and $\overline{E}_{norm}$ is the average power in the answer normalized by the median power for this speaker. We defined a "lively" answer to be one whose liveliness, by this metric, was over 3.5. This can be considered to be a special case of the general correlation between "happiness" and high volume and high pitch (Murray and Arnott, 1993).

We then looked in the corpus to determine how the tutor responded to lively answers. We initially expected these responses to be prosodically somehow different, perhaps being "lively" themselves, as the judge suggested, but this turned out not to be true in general. Instead, the responses, as seen in Table 3, turned out to be notable mostly for the use of unusual acknowledgements. The items *unun*, *haihai*, *soso*, *pinpo-n*[6], *so-da*, and *ye-* (in bold in the table) account for 2/3 of the acknowledgments of lively responses, but only 1/7 of acknowledgements in general.

[6] *pinpon* is a mimetic sound resembling the chime given for right answers on game shows.

order	'liveliness' of answer	corresponding response
1	4.3	**soso**
2	4.1	so
3	3.9	hai
4	3.9	**ye-**
5	3.9	**pinpo-n**
6	3.8	**pinpo-n**
7	3.7	**haihai**
8	3.7	**sososo**
9	3.6	**pinpo-n**
10	3.6	**ununun**
11	3.5	so
12	3.5	**so-da**
13	3.5	(<keep silent>)
14	3.5	hai
15	3.5	**sososo**

Table 3: Responses to Lively Answers

Since we were not sure which of these responses to use in which circumstance, we decided to use the one closest to the response the system would have chosen otherwise. Specifically, if the other rules suggested the use of *hai*, in a lively context the system instead used *haihai*, and similarly it replaces *so*, *un*, and <echo> with *soso*, *unun*, and *so-*<*echo*>, respectively, in response to lively answers.

It is possible to understand this rule as implementing a form of "emotional contagion", that is, the tendency for conversants "to 'catch' each others' emotions, moment to moment" (Hatfield et al., 1994). Specifically, this rule can be seen as implementing the tendency for the tutor to join in the feeling of pleasure in cases where the student is pleased at having found a right answer.

4.3 Rule c: Do Not Wantonly Vary Responses

The preliminary rule set attempted to choose the best acknowledgement in each case, considering only the features of the student's recent utterances. However a judge pointed out that it was unnatural to change acknowledgment response type when things were going smoothly. For example, the following synthesized example sounded inappropriate.

S: SHINBASHI	YURAKUCHO	TOKYO	
T:	un	un	*hai

The system avoids such cases by considering its own previous acknowledgement: if the answer was not preceded by any hints nor incorrect answers, and the time required to answer was shorter

than 1 sec, then the system outputs the same acknowledgement as last time.

4.4 Rule d: Avoid Monotony

On the other hand, a judge also noted that long sequences of the same acknowledgment again and again sounded mechanical and monotonous. In practice, given the other rules of the system, this possibility arose only for a few cases. Thus the system incorporates a rule to switch to *un* after three successive uses of *hai*, and to switch to *hai* after three successive uses of *un*, and similarly for the set of lively responses: *un-un*, *hai-hai*, and *so-so*.

4.5 Rule e: Be Patient when the User is Having Difficulty

Taking a long time to answer is an indication that the student is having trouble recalling the station name. This correlates highly with lack of confidence in the answer, in Japanese as in English (Kimble and Seidel, 1991; Brennan and Williams, 1995).

In the corpus there were 15 cases where the student took more than 30 seconds to answer, and common responses in this situation were: <echo> (5 cases), *pinpo-n* (5), *so-da* (2). Here is an example.

S: (7 second pause)	(1.5 second pause) GOTANDA	
T:	suujideiuto 5	GOTANDA
S: (7 second pause)	(1.5 second pause) GOTANDA	
T:	there is a five (go) in the name	GOTANDA

The system therefore echos back the station name in such cases, with the threshold at 12 seconds. An echo serves as an explicit confirmation, appropriate as a response when the student is uncertain. Subjectively it also seems to indicate 'patience': the willingness to go slow and proceed at the student's pace. It sounds more kindly and less rushed than the default *hai*.

Incidentally, the time-to-response was measured from the end of the system's previous confirmation to the start of the answer. This was a little bit inaccurate in cases where the student's response was preceded by a filled pause, as in *eeto-Shinbashi* (umm-Shinbashi), but such cases were rare.

4.6 Rule f: Praise After Effort

A judge suggested that the tutor should praise the student when he answered after getting hints. This suggestion makes sense since the need to use hints represents a metric of difficulty, and success on difficult problems of course merits praise.

In the corpus, responses to answers after one or more hints are: *un* (9 instances), <echo> (9), *pinpo-n* (7), *so* (5), *hai* (5), *soso* (4). Although *so* is not the most frequently item, it

S:	NISHINIPPORI
T: hora NIPPORI ni chikaindayo	so
S:	NISHINIPPORI
T: similar to NIPPORI	right

is typically found in these contexts: more than half (9/16) of the occurrences of *so* and *soso* appear after answers after hints. Thus, we decided to use *so* for this case. Cut-and-paste conversations generated using this rule do sound better, although we are not sure that whether this is exactly implementing the judges suggestion. There are at least two alternative (or perhaps complementary) explanations. One is that, if the tutor is giving hints, his acknowledgements should not just objectively report correct/incorrect, but also reflect his satisfaction at finding that his hints were useful. Another is that, by giving a hint, the tutor is introducing a referent, and the use of *so* indicates that the student's answer matches that mental object.

There were two sub-cases, depending on whether the user seemed ready to go on, or whether he seemed to still be in a rough patch. In the former case (f1), specifically when the answer took less than two seconds, the system produced a crisp *soso*[7]. In the latter case (f2), the system produces a simple *so*, which seems to encourage the user to take his time coming up with his next response. Also, for the sake of continuity (c.f. Rule c), if the previous acknowledgement was an <echo>, the system produces a *so*-<echo> (f3).

4.7 Rule g: Be Friendly when the User is Uncertain

In the cut-and-paste dialogs, some judges felt some occurrences of *hai* were too "cold". These occurred as responses to answers that required no hints but were uttered without confidence. In these situations the judges preferred *un*. The system therefore outputs *un* in cases where the user has low confidence, as indicated by a non-falling (generally question-like) intonation (Brennan and Williams, 1995).

S: NIPPORI?	
T:	un
S: NIPPORI?	
T:	yes

This rule is subtly different from Rule e. In that case the user has taken a long time to answer and an <echo> is appropriate as a sort of courtesy. In this case, however, the user's intonation is effectively demanding a direct yes/no acknowledgment. Of the two direct confirmations, *hai* and *un*, the latter sounds more appropriate in these cases. This can be explained by saying that

[7] in general, multiple-syllable variants seem to indicate that "I have no more to say, it's your turn, please go on" (Ward, 1998; Ward, 2002).

Rule(s)	A: Condition	D: System Output
a	user is continuing to talk	omit acknowledgement
h (c,d) (b)	no recent incorrect guesses, no hints from tutor	*un* or *hai*
e (b)	user takes more than 12 seconds to produce a guess	<echo-station-name>
g	one or no hints from tutor, rising final intonation (pitch slope greater than 10% per second)	*un*
f1	after a hint or a wrong guess; less than 2 seconds of silence before guess	*soso* or *so*-<echo-station-name>
f2 (f3/b)	after a hint or a wrong guess; more than 2 seconds of silence before guess	*so* or *so*-<echo-station-name>
e (b)	user takes more than 1.5 seconds to produce a guess	<echo-station-name>
g	final pitch not falling (pitch slope greater than –2% per second)	*un*
b (c, d)	pitch and/or energy higher than average (average_pitch_level_in_guess / global_avg._pitch + 1.5×avg._energy_in_guess / global_avg._speech_energy > 3.5)	*un-un*, *hai-hai* or *so-so*
h (c, d)	default	*hai*

Table 4: Response Rules for a system able to respond to subtle, fleeting changes in the user's internal state in the Yamate Loop quiz domain. The D column shows the acknowledgement produced by the system in each condition. The A column specifies the conditions, as determined by the recent context (how many hints the tutor has given the user, how many wrong guesses he has made, how long he has been silent) and by the prosody (pitch and energy contours) of his utterance. Note that "the answer is correct" is implicit in each condition.

a rising intonation shows low confidence and a feeling of dependency, to which the tutor should respond by becoming less formal and more involved or friendly.

4.8 Rule h: Use *hai* as a Default

As a default response, we decided to use *hai* in case none of above rules were applicable, because *hai* is by far the most common acknowledgment in the corpus.

4.9 The Rules as a Set

Although we have independently motivated and presented each rule, in fact they function as a set. Thus it is crucial to adjust the various parameters and organize the rules to work well together. To keep things simple the basic architecture is a list, in which the conditions of each rule are checked in order: the first rule that applied determines which acknowledgement to use.

However there were some complications to this basic method. First, Rules e and g were each implemented in two parts: in an early stage the system checked if Rule e or Rule g was clearly

Rule(s)	B: User's Inferred State Internal	C: System's Internal State in Response
a	unusually confident (rapid pace)	passive
h (c, d) (b)	confident	normal
e (b)	struggling but not wanting help	backing off, slowing down the pace of interaction
g	struggling and wanting help or reassurance	involved, informal
f1	regaining confidence	praising the user, signaling "back on track"
f2 (b)	unsure and wanting support	praising the user for a difficult success
b (c, d)	pleased with him/herself, lively	pleased with the user, excited
h (c, d)	neutral	businesslike, formal

Table 5: Interpretations for the Rules in Table 4

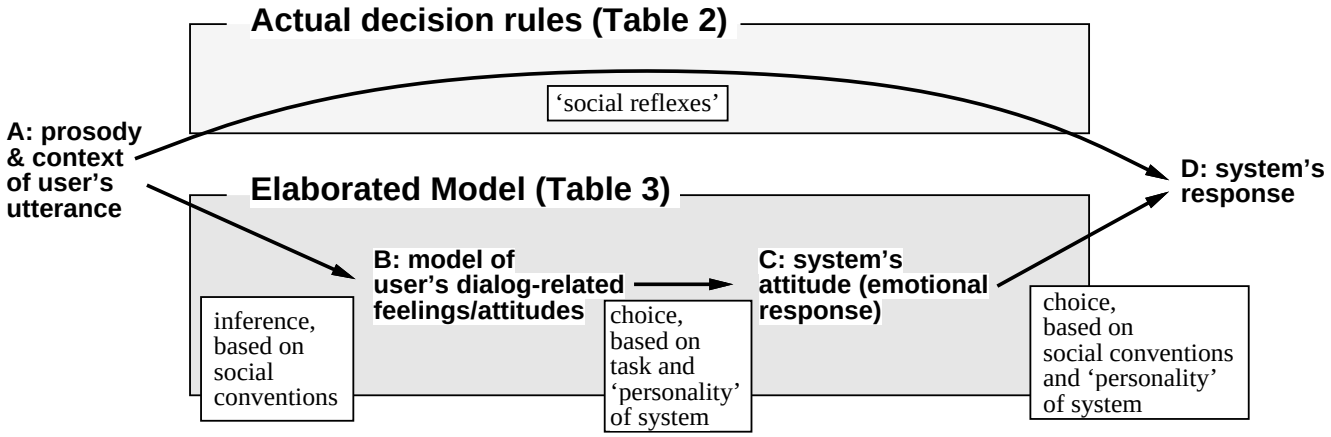

Figure 3: Figure: Two Architectures for Responding to the User's Internal States. A, B, C, and D refer to the columns in Tables 4 and 5.

appropriate, and later, if no other rules applied, it checked them again using weaker conditions. Second, Rule b (use a livelier acknowledgement) operates orthogonally to some of the other rules, thus its conditions are checked independently. Third, Rules c and d (be consistent but not monotonous) are also orthogonal to the other rules, and their conditions are also checked independently.

The rules are summarized in Table 4. Implementation details are given elsewhere (Tsukahara, 2000).

4.10 Interpreting the Rules

There are two ways to look at these rules.

The first viewpoint sees them as pure reflex behaviors, implementing fixed social conventions,

with no deeper significance. This is the view we preferred when we started this project: we sought to build a simple reactive system, inspired by arguments that appropriate social behavior can be explained and implemented without use of inference about the other's internal state, and also without implementing any internal state for the agent (Brooks, 1991; Fridlund, 1997; Ward, 1997). We thus ended up with rules describing direct mappings from prosodic and contextual properties of the subject's guess to the system's response, as seen in Table 4.

A second viewpoint, that these rules embody inferences about the user's attitudes and feelings, was something that we came to later. In a sense this perspective was forced upon us: when we attempted to summarize judges' and subjects' comments about the system, and when we tried to explain the system's rules to other people, it became necessary to add such interpretations. Thus the rules relate to the user's internal state, as shown in column B of Table 5, and to the 'feelings or attitudes' which this system takes in response, as shown in column C of Table 5. We stress that these interpretations are post hoc and tentative (Section 8).

Of course these two viewpoints are not incompatible. Figure 3 suggests the relation between them. The upper path, directly linking the system's inputs and outputs, corresponds to Table 4 and our actual implementation. The bottom path corresponds to the elaborated account in Table 5.

5 EVALUATION METHOD

Our hypothesis is that users prefer a system which is more responsive to their attitudes and feelings and/or the signals which indicate these.

Having developed the rules carefully, paying attention to the corpus and the literature and to judges opinions, we were fairly confident that users would like the system. As a check, we synthesized more dialogs (by audio cut-and-paste) using the final rule set, and played them to naive judges: as expected, acknowledgments chosen by the algorithm generally were rated more natural than conversations with the acknowledgements chosen randomly (Tsukahara, 1998).

The evaluation was thus to determine whether sensitive responses would be preferred by people who were actually interacting with the system, rather than just observing its outputs.

5.1 Control Condition

To determine if sensitively chosen responses were worthwhile, the choice of control condition is critical. The obvious control condition would be to respond invariably with *hai*, the most frequent, neutral, and polite acknowledgement. However in the course of developing the rules it became clear that there is a strong 'variation preference': most people strongly dislike monotonous responses, considering them cold and formal. (Also in human-human interaction, high lexical diversity leads subjects to judge the speaker as having high communicator competence and effectiveness (Berger and Bradac, 1982).) The control condition chosen was therefore a version which chose acknowledgements at random, while preserving the frequencies of each acknowledgment in the corpus. This is a fair baseline, since it has variety and indeed exhibits the

full behavioral repertoire of the system, and so should be as impressive as any system which does not actually use information about the user's state.

5.2 Implementation

Acknowledgements are of course meaningless in isolation, so we built a full system to allow subjects to engage in the Yamate Loop game, providing them not only with acknowledgements but also with appropriate hints. Overall, we wanted to make the experience as similar to a human-human dialog as possible. This constrained the set-up in several ways.

One implementation issue that arose was that of recognition errors. Pilot studies revealed, not surprisingly, that users are very sensitive to mis-recognitions: if a system incorrectly treats a user's guess as wrong even once, that dominates the user's impression of the system as a whole, completely masking the effects of acknowledgement choice. We therefore used a Wizard of Oz set-up, where the experimenter listens to the user's guesses and presses the `y` key or the `n` key, depending on whether the guess was correct. Everything else, including chosing the timing at which to respond, extracting the prosody of the guess, chosing a hint of appropriate easiness, and of course chosing and outputting the acknowledgements, is done by the system.

A second implementation issue was that of speed. In fast-paced dialogs like these, the window of opportunity for a response to be relevant is fairly narrow. We guess that this is on the order of a second or two, based on the casual observation that people in conversation who consistently fail to respond within this time frame appear to be inattentive or socially incompetent or both. To be on the safe side, we made the system able to respond to the user's utterances at the same swift pace as the model human tutor did. In particular, acknowledgements were produced at slowest 360 milliseconds after the end of the speaker's utterance. This was possible because we used a wizard instead of speech recognition and because the wizard practiced until he was able to classify most inputs before the user had finished the word, although at the cost of occasional errors. Thus the user was never kept waiting; this allowed the dialog to continue at a cycle time of as little of 1.6 seconds from one guess to the next. Indeed the pace of interaction was so swift that most users got completely involved in the game of recalling as many station names as they could.

A third implementation issue was that of handling out-of-task utterances. In the corpus there were several cases where one of the participants broke the simple guess-confirm routine with a comment, joke or other digression. Rather than deal with these, we limited the system runs to 90 seconds; this proved to be short enough that digressions did not occur.

A fourth issue was of making the acknowledgements sound natural. For this, we used pre-recorded speech samples, rather than text-to-speech output.

A fifth issue was that of controlling the prosody of the acknowledgements. It was clear from judges opinions of the cut-and-paste dialogs that acknowledgements with inappropriate prosody were immediately noticed and strongly disliked. Since we had decided not to address prosody in this work, we decided to use acknowledgements recorded in a fairly neutral prosody, and these seemed to be generally acceptable.

5.3 The Solicitation of Ratings

Before building the system for the live experiments, we did a small preliminary study with people listening to cut-and-paste synthesized conversations. Initially we asked people to judge the quality of the system's contribution, without giving them any hint of what to pay attention to. Subjects seemed to find this difficult, and there was no clear preference for either system. We then tried again, this time telling the subjects that this was a experiment about acknowledgement and asking them to judge the conversation mainly focusing on the tutor's responses. This gave clearer results.

For the live experiments it would have been simpler had we been able to elicit preference judgements without calling subjects' attention to the acknowledgments, but based on the prelimary study we decided that doing so was probably necessary, for several reasons.

First, the differences between the system acknowledgements probably fell below the level of conscious awareness for most subjects. This seemed likely to be even more of a problem for the live experiments, where the game was so fast-paced that users were likely to be too busy trying to recall station names to have any attention to spare to think about the systems responses. This problem would of course not arise in most spoken dialog systems, where the the pace of interaction is much slower, and users are left with free time to contemplate the prompts and responses of the system. It would also not have arisen had the system produced full sentences like *good, at last you got it, please keep it up, beep*, where inappropriate acknowledgements would be much more salient. It also would not have arisen were the baseline not so high; both our system and the control system were operating at near-human levels of performance, with no gross infelicities.

Second, memories fade, especially regarding short-lived attitudes and feelings, whether one's own or those projected by the dialog partner. By the the end of listening to, or engaging in, a dialog, users probably don't remember how they felt at each moment during the interaction. Even if they were momentarily irritated or confused or amused or pleased by the system at various times, but after a minute or two, when asked "how did you rate the system", those impressions have probably been forgotten.

Third, each dialog was short, only 90 seconds. With longer dialogs or extended use, say over 5 to 10 minutes, there would probably be clearer effects: the cumulative effects of minor awkward choices would probably accumulate and create an poor overall impression, or conversely the cumulative effects of consistently saying just the right thing would lead to an overall impression of high-quality.

In response to these problems, we considered several ways to get subjects to pay sufficient attention to the acknowledgements. We considered telling users up front to pay attention to the acknowledgements, but this probably would have changed their behavior. We considered ways to get evaluations from users as they were interacting with the system, such as having them think aloud, or pausing the interaction after each acknowledgement (Teague et al., 1991), but this would have destroyed the real time nature of the interaction. We considered using third-party observers, to watch and listen to the subjects interacting with the system, either live or recorded, and to judge the quality of the interaction, but it is known that third-party

observers' opinions of what constitutes a good interaction do not always agree with the opinions of participants.

Finally we chose to use 'retrospective testing' (Neilsen, 1993). This is a standard interface evaluation method in which the user views or listens to a recording of his interaction, and reports how he felt during each moment of the interaction. In a sense, this technique allows the amplification of weakly-detected user preferences, by allowing the user to devote full attention to the task of evaluating system quality, while at the same time getting access to the user's private knowledge. Of course, it is impossible to know for sure whether users are reporting their true affective states at each point or reporting on how they think they were *supposed* to have felt at each point in the interaction[8]. Certainly there is no guarantee that these judgements are accurate, but on the other hand there is no particular reason to think that users have internalized social norms or naive beliefs about how people are supposed to behave or feel in such tasks. Thus, overall, retrospective testing seemed to be the least problematic of the possible evaluation methods. Nevertheless, we performed various cross-checks, as described below (Section 6.2); these suggested that the judgements after re-listening were indeed more accurate than the initial judgements

During re-listening we allowed the user to stop or rewind the play-back at will. To make evaluation easier, and to draw attention to the acknowledgements, the user was given a transcript of his interaction with the system, with a tiny 7-point scale printed above each of the system's acknowledgements, for him to mark his rating. In order to have this transcript available immediately, before the users' impressions could fade, it was computer-generated and automatically sent to the printer after the session ended.

5.4 The Protocol and the Subjects

The subjects were juniors participating in experiments to fulfill a class requirement. Before the experiment itself, we had them read aloud an unrelated list of station names, so the system to measure their normal pitch and volume levels and ranges.

Each user interacted with the full system and the random version for about 90 seconds each. The order of presentation was chosen at random. The two runs covered different segments of the Yamate Loop line. Subjects were requested just to "use this system".

All subjects found this a reasonable task and were able to interact with the two versions. Most users believed they were interacting with a fully automatic system, and yet their behavior was, it seemed, as natural as if they were talking to a human.

Subjects were excluded from the analysis in cases where the wizard misclassified an utterance, where there were less than three acknowledgements in each run, or where the number of acknowledgements occurring in the two runs differed by a factor of two or more, which happened typically when the user was less familiar with the station names in one segment of the Yamate Line. In the end there remained usable data from 13 subjects.

After interacting with each system, subjects answered two questions: "Which computer

[8] We thank an anonymous reviewer for pointing this out this problem.

would you like to use?" After this we told them we were interested in acknowledgements, and asked "How would you rate the overall naturalness of the acknowledgements produced by the system?": this gave us their first impressions[9]

Then they listened to their dialogs, ranking the naturalness of each acknowledgement, on a 7 point scale. After this was complete, we had them rate the two systems on various dimensions, such as naturalness, friendliness, and patience, again using 7-point scales. After this we asked again: "Which computer would you like to use?"

Finally, we told the users the purpose of the experiment and asked them for comments and suggestions.

6 RESULTS

6.1 Main Result

10 out of 13 subjects preferred the system which produced acknowledgements by rule to the one that produced them randomly ($p < 0.05$ by the sign test).

While this result is only just significant, it is corroborated by a preliminary experiment, identical in all respects other than that the hints were produced by the wizard, not automatically, in which 12 of 15 users preferred the system that did rule-based choice of acknowledgements. The rule-based system was also significantly better in that the individual acknowledgements were generally ranked higher on the 'naturalness' dimension for the rule-based system ($p < 0.05$ by the U-test, 7-point scale). User comments generally also were compatible with the interpretation that the system which chose acknowledgements by rule was better (Tsukahara, 2000).

6.2 The Value of the Re-Listening Phase

Regarding the question of whether the judgements after re-listening are "better" than first impressions, unfortunately there is no direct evidence; only some incidental indications that re-listening helps users make more informed judgements:

First, users' preferences were clearer after re-listening, with fewer cases of "indifferent" or "no preference".

Second, preferences were more internally consistent. For one thing the judgements of the appropriateness of specific responses correlated better with judgements of the usability of the system as a whole. For another overall judgements on the various scales were more consistent. For example, before re-listening, 6 subjects' preferred system was not the one which they rated the most kind, but after re-listening no subjects' ratings had these contradictions.

[9] In retrospect, it would have been good to have also had users evaluate the system in terms of understanding, rapport, reinforcement, and supportiveness, factors that are known to be important in judgements of attractiveness of people (Berger and Bradac, 1982).

Third, after re-listening, subjects volunteered more comments regarding the appropriateness of individual items (before re-listening 4/13 subjects gave comments and after re-listening 11/13 gave comments, $p < 0.05$).

Fourth, regarding user's preferences at first impression, there was a slight tendency for users to prefer whichever system they used second, presumably due to a gain in familiarity with the task. This tendency disappeared after re-listening.

Fifth, the results obtained after re-listening were more consistent with the results of the evaluations where judges listened to synthesized dialogs. This of course raises the question of, in general, when it is necessary to do live experiments, and when it suffices to rely on the opinions of third-party judges. This is a question requiring further research.

6.3 Individual Differences

The fact that there were subjects who did not prefer the more responsive system is also interesting. In part this was probably due to uninteresting factors: chance, bugs in the rules used in the current system, failures to vary the prosody of each acknowledgement, etc. But we suspect that there also are individual differences in the style of interaction preferred by users. One of the users who preferred the random system commented that "when it confirmed by repeating the station I had just said, it felt fake, like it suddenly had gotten perfectly in touch with me"; perhaps this user would have preferred a more mechanical, formal, style of interaction. This effect was also seen in the tendency (not significant) for the random-preferring subjects to rank the acknowledgements produced by the "default" rule (*hai*) more highly than other acknowledgements; whereas for the rule-preferring subjects the opposite was generally the case. Maybe personality traits, such as introversion/extroversion or reactions to being monitored and thoughts about personal control (Bickmore and Cassell, 2001; Rickenberg and Reeves, 2000), are involved here, which of course raises the question of how to detect and adapt to the different interactional styles and preferences of users.

6.4 Miscellaneous Observations

Regarding the contributions of each of the individual rules to the user's preferences, the effects of the "omit acknowledgements" rule (Rule a) were generally rated poorly, however we believe that this was due to a poor choice of parameters, rather than a mistake with the rule itself. Regarding the other rules, no significant results were found (Tsukahara, 2000).

Regarding the exact nature of the system's responsiveness, in designing the system our thought was that it would be the exactly appropriate acknowledgements, even if only sometimes present, that would be of value to the user. However some of the user's comments seem to suggest that the advantage of the rule-based system lies elsewhere: in its ability to provide variety while avoiding the crashingly-bad acknowledgements which random choice occasionally gives. Further study is needed.

Regarding the generality of the finding across languages, we suspect that similar results could be obtained for systems in other languages. Certainly the observations seen in Table 1 are not

unique to Japanese. The generality may be even stronger, however. At CHI 2001 in Seattle we illustrated the system with two videotaped 90 second segments, and the audience's show of hands was two dozen to one, preferring the system producing rule-based acknowledgements to the random one, even though the dialogs shown were in Japanese. This suggests that there may be properties of responsiveness in dialog which are true across languages (Ward, 2002).

7 IMPLICATIONS

Responsiveness in interfaces is a fairly new research topic, but one that relates to two strong themes in interface research: the dream of systems which infer the user's implicit emotional state, and the dream of systems which follow the conventions of social interaction.

7.1 On Emotions in Interfaces

The first dream is that of exploiting emotions in interfaces. If we consider the system to be modeling the user's emotional state (admittedly a problematic assumption Section 8), then the current finding is perhaps the first demonstration that emotion modeling can actually have value for the user. This sub-section speculates why this might be true.

Emotional interfaces (Picard, 1997; Ball and Breese, 2000; Cowie et al., 2001) are sometimes seen as an antidote to the coldness of the 'purely rational' interfaces common today. Given this antithesis, it is natural that most attention has focused on the 'classic' emotions, such as joy, anger, sadness, arousal, and fear. However, no compelling need for the detection of such user emotions has yet been identified, as users are not generally emotional in these ways. On the system side, giving such emotional states to systems of clear value for entertainment purposes (Bates, 1994), but it is not clear whether this is worthwhile for systems which users need to interact with, rather than just watch. (Shinozaki and Abe's (1998) tutorial system is not an exception: although the version of the system with more moods was rated more highly, the subjects' task was not to actually remember or learn anything, but rather to use the system to "experience various instruction messages".) We believe that the focus should instead be attitudes and feelings of the sort seen in Table 1 because, if the goal is to improve user interfaces, it makes sense to use those emotions which are the most common and the most related to communication and social interaction.

One issue in emotionally-aware interfaces, and in user modeling more generally, is the double-edged danger: on the one hand that failure to read the user's emotions accurately may annoy him, but on the other that success may make him feel deprived of control, especially if he has no clue why the system reacted as it did (Lanier, 1995). The sort of emotions discussed above are however less prone to this danger. Feelings and attitudes that relate to the current state of the interaction can be occasionally mis-understood with no persistent effects, since the effects of a single bad inference do not affect subsequent exchanges. That is, in case of failure, the system may seem momentarily cold, out-of-synch, non-attentive, foreign, or perhaps robotic, but in a content-dominant situation the system will have a chance to redeem itself in the next moment. Regarding success, with dialog-related attitudes and feelings this need not belittle the user, as

the underlying assumption is not that the user is incapable of expressing what he or she wants, but that the user is clearly (albeit non-verbally) indicating his or her feelings and needs.

Another issue in emotional interface research is the sheer complexity of inferring and responding to emotions. The attitudes and feelings addressed in this paper are much easier to deal with than the classical emotions. Since the responses they evoke come so swiftly, users don't expect anything beyond simple reflex-type responses. (There are limits to how much people are able to process in a fraction of a second, and systems need do no more than this.) Thus users do not expect the sort of powerful inference (for the sake of inferring the user's intention or knowledge from indirect or confused statements) of the sort addressed by the user modeling work in the AI tradition. In other words, dealing with dialog-related attitudes and feelings allows a form of user-adaptive behavior that is swift, and that it relies not on careful reasoning or deep domain knowledge, but rather on simple features of the context and on non-verbal cues provided by the user.

7.2 On Agents which Follow Social Conventions

The second dream is that of building systems which obey social conventions, and especially non-verbal and real-time conventions (Cassell et al., 2000). In the long term, dialog systems which are unable to handle social conventions seem destined to have only limited user acceptance (Johnstone et al., 1995).

Recent work on real-time social conventions includes mostly work on turn-taking: the process by which two speakers smoothly take turns, without awkward silences or talking over each other, and without explicit protocols ("roger, over and out"). Schmandt (Schmandt, 1994) built a system which gave driving directions and used the length and pitch slope of user utterances to control the pace of its delivery. Thorisson and Cassell's (1999) Ymir was a multi-modal animated system which detected the onset and offset of the user's voice, among other things, and used this to determine when to be listening/not-listening and taking-a-turn/yielding-the-turn; the version of the system which did this was ranked higher and considered more "helpful" by users. Ward and Tsukahara (1999) built a system which detected a prosodic feature cuing back-channel feedback (*uh-huh* etc.) and responded appropriately. Cassell *et al.*'s (1999) Rea used several types of information (user present/absent, user speaking/silent, declarative/interrogative/imperative user's utterance, user gesturing/still) to determine when the agent should perform various actions.

However, this current work is probably the first to show that obeying real-time social conventions is actually preferred by users. Previous evaluations have been complicated by two factors. First there is the Eliza effect: that users tend to cooperatively ascribe sense to what the system does, regardless of whether it is appropriate or just random variation. Second, there is the 'variation preference': the basic human preferences for characters that are more active and exhibit a wider repertoire of actions. In the current work the reference system was chosen to control for the variance preference, thus we can conclude that users really do prefer systems that accurately implement human social interactional conventions. That is, our results show that there is indeed a payoff for work in this research area: using social conventions can result in a system which measurably improves the user experience.

Regarding the use of social conventions in interfaces, there are serious doubts about the value of these (Shneiderman, 2000), or of anthropomorphism more generally. However these doubts are probably less relevant to *real-time* social conventions. For one thing, the problem that doing so can make system behavior unpredictable for users is not a great problem when the behavior is orthogonal to the content of the interaction, having no effects on the downstream behavior of the system. For another thing, real-time conventions generally involve a separate channel (or are 'out-of-band' or 'in a separate modality') from the main interaction channel, and so incorporating them is probably less likely to interrupt or distract users. (In some respects short acknowledgements probably function as a separate channel (Jaffe, 1978; Goffman, 1981; Ward, 2002).)

8 PROSPECTS AND OPEN QUESTIONS

We foresee that the sensitive modeling of user state in dialogs will find applications first in simple tutorial systems: for example a system to assist multiplication table memorization, perhaps made available over the telephone via a 900 number. If users prefer a more responsive system over a less responsive one, they may use it more, and learn more, if only due to increased time-on-task.

Further along, we see this as a value-added component to systems for all sorts of tasks. When spoken dialog system developers begin to automate dialogs which are not merely clerical, but which involve persuading, motivating, charming, and selling, they will need to copy the sensitivity and style of superior human communicators — great teachers, great bartenders, great salesman and great bosses. Two large problems, however, remain.

First there is the basic problem of speech recognition accuracy. Full implementation of a system even for our simple experimental scenario would require two advances: the ability to recognize words in progress, since a system should be able to determine the import of an utterance before the user finishes talking, and the ability to recognize words even when they are stretched or distorted or padded with fillers, as produced by users who speak while they are still thinking.

There is also an architectural issue. If the aim is to build a full agent able to respond to emotions, including dialog-related attitudes and feelings makes the implementation harder. This is not just an algorithmic or hardware problem but also one of design: perhaps requiring multiple simultaneous threads of control (something not supported by today's standard architectures for dialog management) in order to allow reactive (shallow, emotion-based, conventional) responses to execute swiftly and somewhat autonomously from more deliberative, content-based response planning (Cassell and Thorisson, 1999).

Then there is the problem of development cost. This project took 3 man-years: which is clearly not cost-effective for practical purposes. Of course, it would be possible to develop a similar system faster today, by avoiding the dead-ends and pitfalls described in Sections 3 and 4. However what is still lacking is a clear understanding of the roles of feelings and attitudes in dialog. Our inventory of these, in Table 5, was arrived at *post hoc*. There is a need for a systematic analysis and general theory that can serve to guide the designs of future sensitive systems (Cowie et al., 2001).

Given some basic research in this area, it would be possible to develop future systems with much less investment of time. However the design of responsive interfaces will probably never be trivial. This is because the pragmatic force of non-verbal signals is highly task- and context-dependent. For example the parenthesized aspects in Table 1 will depend on the specific task domain and in the worst case will require dynamic inference using the context. Moreover the personality that the system is to project, which determines the B-to-C mappings in Figure 3, will also need to vary from system to system.

Clearly the present study is a preliminary exploration. In particular, there are many ways in which the experimental methods ought to be refined. Here we mention just four.

The first question is that of how valuable it actually is to add this sort of responsiveness. The experiment showed a significant user preference for the more responsive system, but not whether difference was a just-noticeable one, or something of substantial value. Certainly none of our subjects was wildly enthusiastic about the abilities of our system: there was no "wow" effect, of the sort that is anecdotally reported for people interacting with some animated agents (Massaro, 1997). If the goal is merely to create systems which are human-like and "believable", or give a positive first impression, or evoke perceptions of social competence, it suffices to give a system a face, animated actions, the ability to track the user with eyes, or even just an identifying name or color (Lester et al., 1997; Reeves and Nass, 1996), all of which are simpler to implement than subtle sensitivity and responsiveness. However there may be applications where it is worth the added effort to go beyond mere human-ness, to include behaviors that are finely tuned to the user's states and actions in the micro-scale and in real time, as done here.

The second big open question is that of the theoretical status of this work. We have found that it is possible to build a system which is usefully responsive, without being clear about what it is actually doing. Thus, as noted above, the inventory of attitudes and feelings in our system is post hoc, and our initial list of phenomena of interest (Table 1) clearly mixes expressions of emotional state, attitudes, conversational flow control intentions, and social conventions. However, for this line of work to make contact with, and take advantage of, the various insights provided by cognitive psychology, social psychology, and related fields, it would help enormously to regularize the terminology and rigorously categorize the phenemona. Thus, situating this work in some clear theoretical framework is an important problem for further research.

The third big open question is that of the value, indeed the appropriateness, of using re-listening for obtaining user opinions. Although practically it seems useful, its general validity is open to question. Ultimately the question of how to obtain preferences ties to larger questions, such as whether people really know what they like, how outside observers can measure this, what it means to prefer something, and to what extent preferences are stable over time and across different measurement methods.

The fourth big open question involves the nature of the preference for the more responsive system. Responses to our main question "which computer would you like to use" generally correlated with responses to other questions, such ratings of naturalness, friendliness and patience, however we did not have enough subjects to explore this in depth. Certainly preference and ranked "naturalness" etc. will not always correlate. More work needs to be done here.

9 SUMMARY

Building spoken language systems that operate at near-human levels will doubtless require lots of attention to the 'little things' in dialog, individually minor, but in aggregate determining whether users find the system to be fun to use or just tolerable. We have identified some of these "little things".

We have identified the role of some dialog-related attitudes and feelings in human interaction, argued that they can be important in real-time interactive systems, and verified this by experiment. This result shows also that using real-time social conventions can result in a system which measurably improves the user experience, if a sufficiently sensitive measure is used. Methodological lessons learned include the value of modeling the system on the behavior patterns of a single individual, and the value of evaluation after re-listening as a way to sharpen judgements of usability. Exploiting these findings will be difficult, however, without advances in the recognition of words uttered during thought, advances in the identification of prosodic and non-verbal cues in dialog, and advances in the understanding of attitudes, feelings, and emotions and their role in human interaction. Nevertheless, we see this line of work as ultimately essential to the development of truly effective interface agents, able to persuade, motivate, charm, and sell.

10 ACKNOWLEDGMENTS

We thank the Nakayama Foundation, the Inamori Foundation, the International Communication Foundation, and the Japanese Ministry of Education for support, our subjects for their cooperation, and two anonymous reviewers for comments and suggestions. This work was done while the first author was at the University of Tokyo.

References

Angles, J., Nagatomi, A., and Nakayama, M. (2000). Japanese responses *hai, ee*, and *un*: yes, no, and beyond. *Language and Communication*, 20:55–86.

Ball, G. and Breese, J. (2000). Emotion and personality in a conversational agent. In *Embodied Conversational Agents*, pages 189–219. MIT Press.

Bates, J. (1994). The role of emotion in believable agents. *Communications of the ACM*, 37:122–125.

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

Berger, C. R. and Bradac, J. J. (1982). *Language and Social Knowledge: The social psychology of language*. Edward Arnold, Ltd.

Bickmore, T. and Cassell, J. (2001). Relational agents: A model and implementation of building user trust. In *SIGCHI'01*, pages 396–403. ACM.

Brennan, S. E. and Williams, M. (1995). The feeling of another's knowing: Prosody and filled pauses as cues to listeners about the metacognitive states of speakers. *Journal of Memory and Language*, 34:383–398.

Brooks, R. A. (1991). Intelligence without representation. *Artificial Intelligence*, 47:139–159.

Cassell, J., Bickmore, T., et al. (1999). Embodiment in conversational interfaces: Rea. In *CHI '99*, pages 520–527. ACM Press.

Cassell, J., Bickmore, T., et al. (2000). Human conversation as a system framework. In *Embodied Conversational Agents*, pages 29–63. MIT Press.

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

Clark, H. H. (1996). *Using Language*. Cambridge University Press.

Cowie, R., Douglas-Cowie, E., and Romano, A. (1999). Changing emotional tone in dialog and its prosodic correlates. In *Proceedings of ESCA International Workshop on Dialog and Prosody*, pages 41–46.

Cowie, R., Douglas-Cowie, E., Tsapatsoulis, N., Votsis, G., Kollais, S., Fellenz, W., and Taylor, J. G. (2001). Emotion recognition in human-computer interaction. *IEEE Signal Processing Magazine*, 18:32–80.

Fridlund, A. J. (1997). The new ethology of human facial expressions. In Russell, J. A. and Dols, J. F., editors, *The Psychology of Facial Expression*, pages 103–129. Cambridge.

Goffman, E. (1981). Response cries. In Goffman, E., editor, *Forms of Talk*, pages 78–122. Blackwell. originally in *Language* 54 (1978), pp. 787–815.

Graesser, A. C., Wiemer-Hastings, K., et al. (1999). Autotutor: A simulation of a human tutor. *Journal of Cognitive Systems Research*, 1:35–51.

Hatfield, E., Cacioppo, J. T., and Rapson, R. L. (1994). *Emotional Contagion*. Cambridge University Press.

Jaffe, J. (1978). Parliamentary procedure and the brain. In Siegman, A. W. and Feldstein, S., editors, *Nonverbal Behavior and Communication*, pages 55–66. Lawrence Erlbaum Associates.

Johnstone, A., Berry, U., Nguyen, T., and Asper, A. (1995). There was a long pause: Influencing turn-taking behaviour in human-human and human-computer dialogs. *Int. J. Human-Computer Studies*, 42:383–411.

Kimble, C. E. and Seidel, S. D. (1991). Vocal signs of confidence. *Journal of Nonverbal Behavior*, 15:99–105.

Lanier, J. (1995). Agents of alienation. *Interactions*, 2:66–72. also at http://www.well.com/user/jaron/agentalien.html.

Lester, J. C., Converse, S. A., et al. (1997). The persona effect: Affective impact of animated pedagogical agents. In *CHI '97*, pages 359–366. ACM Press.

Mane, A., Boyce, S., Karis, D., and Yankelovich, N. (1996). Designing the user interface for speech recognition applications. *SigCHI Bulletin*, 28:29–34.

Massaro, D. W. (1997). *Perceiving Talking Faces*. MIT Press.

Murray, I. R. and Arnott, J. L. (1993). Toward the simulation of emotion in synthetic speech: A review of the literature on human vocal emotion. *Journal of the Acoustic Society of America*, 93:1097–1108.

Neilsen, J. (1993). *Usability Engineering*. Morgan Kaufmann.

Oviatt, S. (1996). User-centered modeling for spoken language and multimodal interfaces. *IEEE Multimedia*, pages 26–35.

Picard, R. (1997). *Affective Computing*. MIT Press.

Rajan, S., Craig, S. D., Gholson, B., Person, N. K., and Graesser, A. C. (2001). Autotutor: Incorporating back-channel feedback and other human-like converational behaviors into an intelligent tutoring system. *International Journal of Speech Technology*, 4:117–126.

Reeves, B. and Nass, C. (1996). *The Media Equation*. CSLI and Cambridge.

Rickenberg, R. and Reeves, B. (2000). The effects of animated characters on anxiety, task performance, and evaluations of user interfaces. In *CHI '00*, pages 49–56. ACM Press.

Schmandt, C. (1994). *Computers and Communication*. Van Nostrand Reinhold.

Shinozaki, T. and Abe, M. (1998). Development of CAI system employing synthesized speech responses. In *International Conference on Spoken Language Processing*, pages 2855–2858.

Shneiderman, B. (2000). The limits of speech recognition. *Communications of the ACM*, 43:63–65.

Teague, R., De Jesus, K., and Nunes-Ueno, M. (1991). Concurrent vs. post-task usability test ratings. In *CHI 2001 Extended Abstracts*, pages 289–290.

Tsukahara, W. (1998). An algorithm for choosing Japanese acknowledgments using prosodic cues and context. In *International Conference on Spoken Language Processing*, pages 691–694.

Tsukahara, W. (2000). Choice of acknowledgements based on prosody and context in a responsive spoken dialog system (in Japanese). D.Eng. Thesis, University of Tokyo, School of Engineering.

Walker, M. A., Litman, D. J., Kamm, C. A., and Abella, A. (1998). Evaluating spoken dialog agents with paradise: Two case studies. *Computer Speech and Language*, 12:317–348.

Ward, N. (1997). Responsiveness in dialog and priorities for language research. *Systems and Cybernetics*, 28(6):521–533.

Ward, N. (1998). The relationship between sound and meaning in Japanese back-channel grunts. In *Proceedings of the 4th Annual Meeting of the (Japanese) Association for Natural Language Processing*, pages 464–467.

Ward, N. (2000). The challenge of non-lexical speech sounds. In *International Conference on Spoken Language Processing*, pages II: 571–574.

Ward, N. (2002). A model of conversational grunts in American English. submitted to *Cognitive Linguistics*.

Ward, N. and Kuroda, T. (1999). Requirements for a socially aware free-standing agent. In *Proceedings of the Second International Symposium on Humanoid Robots*, pages 108–114.

Ward, N. and Tsukahara, W. (1999). A responsive dialog system. In Wilks, Y., editor, *Machine Conversations*, pages 169–174. Kluwer.

Yankelovich, N., Levow, G.-A., and Marx, M. (1995). Designing SpeechActs: Issues in speech user interfaces. In *CHI '95*, pages 369–376. ACM Press.