Responding to Subtle, Fleeting Changes in the User's Internal State

Wataru Tsukahara
Hitachi Central Research Labs
1-280 Higashi-Koigakubo, Kokubunji-shi
Tokyo 185-8601 Japan
+81-42-323-1111, ext. 3681
w-tsuka@crl.hitachi.co.jp

Nigel Ward
Mech-Info Engineering, University of Tokyo
7-3-1 Hongo, Bunkyo-ku,
Tokyo 113-8656 Japan
+81-3-5841-6346
nigel@sanpo.t.u-tokyo.ac.jp

ABSTRACT
In human-to-human interaction, people sometimes are able to pick up and respond sensitively to the other's internal state as it shifts moment by moment over the course of an exchange. To find out whether such an ability is worthwhile for computer human interfaces, we built a semi-automated tutoring-type spoken dialog system. The system inferred information about the user's 'ephemeral emotions', such as confidence, confusion, pleasure, and dependency, from the prosody of his utterances and the context. It used this information to select the most appropriate acknowledgement form at each moment. In doing so the system was following some of the basic social conventions for real-time interaction. Users rated the system with this ability more highly than a version without.

Keywords
ephemeral emotions, social interaction, spoken dialog, tutoring, real-time, responsive, non-verbal, Japanese, acknowledgements, feedback, prosody

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" [9] of the dialog. Even if there is no real content, as in talk about the weather, or divisive content, such as a difference of opinion, the same feelings of satisfaction can arise. Thus the *process* of dialog itself, not just the result, can be valuable.

Some of this is due, we believe, to successful exchanges of information about the participants' states, in real time as they change moment-by-moment during the dialog.

SIGCHI'01, March 31-April 4, 2001, Seattle, WA, USA.

That is, it can be satisfying when a conversation partner accurately tracks the attitudes and feelings which subtly color most of our utterances. Some of these 'ephemeral emotions' are seen in Table 1.

Dialog participants exchange information about their ephemeral emotional states with various subtle signals, including non-verbal ones, generally sent and received without conscious attention, we believe. However today's spoken dialog systems do not handle this level of communication, focusing almost exclusively on the information content of the interaction. This is true largely because design is currently limited by the need to constrain the interaction in various ways so as to allow the recognizer to perform accurately. Thus 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 [26], and of using various counter-strategies for achieving relatively natural interaction despite these constraints [31, 17]. Looking to the future, however, dialog systems that are blind to such signals will inevitably exhibit the basic "responsiveness blooper" [13] of failing to acknowledge a substantial component of the user's input.

Thus, if we want to build dialog systems that people will find pleasant to interact with, or at least less tiresome, we probably should model these aspects of 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 ephemeral emotional 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.

TUTOR SUPPORT IN MEMORY GAMES

In order to study these phenomena we chose one of the simplest dialog forms we could imagine: practice memory quizzes. If you need to memorize chemical symbols, multiplication tables, geographical names, or the like, to have someone quiz you verbally is a good way to test your knowledge, and to motivate you to study further. Of course, this only works if the tutor is helpful, sympathetic, encouraging,

I want to express my thoughts (by taking a turn soon)	I'm not committed to any opinion (so you're welcome to keep talking)
I'm uncomfortable (with this topic)	I'm bored (so let's talk about something else)
I'm amused (by your story)	I'm concerned (that I'm not expressing myself well enough)
I'm frustrated (that I've not been able to convince you)	I'm really interested (in your opinion on this)
I'm pleased (that you appreciate the irony in my words)	I know just how you feel about that (and sympathize)
I'm missing something (so you need to be more explicit)	I'm aware of that already (so we can go on to talk about something else)
I need a moment (to digest that statement)	I'm getting restless (so let's close out this conversation)
I know what I'm talking about (so please just listen a minute)	I'm feeling a twinge of irritation (at the tone of your last remark)

Table 1: Examples of Feelings that Occur as 'Ephemeral Emotions' 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)

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: Japanese original above; English translation below

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" [11].

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". The other person, playing the role of student, does his or her best to recall them; and after each guess, the tutor tells the student whether he or she was right, and if not, gives hints. There are 27 stations on the Yamate Loop Line; the typical Tokyo-ite knows many but requires hints for the rest. Figure 1 shows a typical dialog fragment. 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, and in particular ephemeral emotions seem to occur, and to be expressed and responded to. More often than not, participants seem to find these little dialogs enjoyable.

Focusing on the tutor's role, we set out to build a system that could take part in such interactions.

ANALYZING REAL-TIME INTERACTION

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. In particular, we focused on word choice, since variation in acknowledgement timing and prosody, although significant, seemed less expressive and less varied, and also seemed harder to analyze. We therefore decided to model the rule by which the tutor chose how to respond to correct station names: specifically how he selected

among the 11 acknowledgement forms seen in column D of Table 2. 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, although there is no simple correspondence between the acknowledgements in the two languages.

In focusing on acknowledgement choice, our work is again similar to that of Graesser *et al.* ([11]). In their system, however, choice among acknowledgements is determined by "the quality of the set of assertions within a conversational turn". Thus their system analyzed the content of the user's input (a text string), not the way the user felt when he or she produced it. We wanted our system to go beyond this, to interpret the user's inputs more 'sensitively'.

We therefore looked for signals from the users which affected acknowledgement choice. We chose to analyze only audio, since, although face-to-face tutoring is best, it is possible to tutor effectively even over the telephone, and because the spoken interaction was all we intended to implement.

To find the signals and rules governing acknowledgement choice we consulted various sources. Dictionaries, unfortunately, have poor coverage of words primarily used in social interaction. There is some research on Japanese acknowledgements, such as [1], but this was too broad-brush to be helpful (we believe the root problem is that linguists like to study rich dialogs, in which too many confounding factors are present to allow the positive identification of any one). There is a large body of work on correlates of emotion in speech signals, but this is almost completely focused on joy, sadness, anger and the other 'basic' emotions, with Brennan and Williams ([4]) being a welcome exception.

We thus had to discover the rules ourselves. Since we had plenty of data, in all more than a thousand acknowledgements, from 41 pairs of people, 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, and 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 have 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. We therefore solicited a further 5 dialogs with him and analyzed these.

Initially we looked for statistical correlations between properties of the input (the context and prosody of the student's correct guess) and the output (the system's acknowledgement). Based on these we hand-coded an initial set of decision rules. 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 highly matched the actual tutor's acknowledgements 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 a 3 friends not familiar with our research. They 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 [23].

RULES FOR RESPONDING TO EPHEMERAL EMOTIONS

The rules we finally settled on and implemented are summarized in Table 2. Implementation details are given elsewhere [24]. 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 without implementing any internal state for the agent [5, 10, 27]. 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 2.

A second viewpoint, that these rules embody inferences about ephemeral emotions, was something that we came to later. In particular, when we reviewed judges' and subjects' comments about the system, and when we tried to explain these rules to other people, it became necessary to add such interpretations. Thus the rules relate to user emotions, as shown in column B of Table 3, and to the 'emotional states or attitudes of the system' taken in response, as shown in column C of Table 3.

Of course these two viewpoints are not incompatible. Figure 2 suggests the relation between them. The upper path, directly linking the system's inputs and outputs, corresponds to Table 2 and our actual implementation. The bottom path corresponds to the elaborated account in Table 3.

In passing we note that, despite the intrinsically asymmetric nature of the interactions in this task, there is still an element of "emotional contagion", that is, the tendency for conversants "to 'catch' each others' emotions, moment to moment" [12], in mapping number 6.

EVALUATION METHOD

Our hypothesis is that users prefer a system which responds appropriately to their ephemeral internal state.

We accordingly ran experiments having subjects interact with two versions of the system: one which used the rules in Table 2, and one which did not. The obvious control con-

A: Condition	D: System Output	Code
user continues talking (immediately uttering the next station name)	omit acknowledgement	1
no recent incorrect guesses, no hints from tutor	*un* or *hai* *	2
user takes more than 12 seconds to produce a guess	[station-name] (echo)	3a
one or no hints from tutor, rising final intonation (pitch slope greater than 10% per second)	*un*	4a
after a hint or a wrong guess; less than 2 seconds of silence before guess	*soso* or *so*-[station-name] *	5a
after a hint or a wrong guess; more than 2 seconds of silence before guess	*so*	5b
user takes more than 1.5 seconds to produce a guess	[station-name] (echo)	3b
final pitch not falling (pitch slope greater than -2% per second)	*un*	4b
pitch and/or energy higher than average (average_pitch_level_in_guess / global_average_pitch + 1.5 $\times$ average_energy_in_guess / global_average_utterance_energy > 3.5)	*un-un*, *hai-hai* or *so-so* *	6
default	*hai*	7

Table 2: 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 middle 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. The "code" column is a cross-reference to Table 3. The conditions are checked in the order shown and the output is determined by first rule whose condition applies. * indicates cases where further sub-rules are applied: these prevent wanton variation in acknowledgement from one time to the next, and yet prevent long monotonous sequences of identical acknowledgements.

Code	B: User's Ephemeral Emotion Inferred	C: System's Ephemeral Emotion in Response
1	unusually confident (rapid pace)	passive
2	confident	normal
3a,b	struggling but not wanting help	backing off, slowing down the pace of interaction
4a,b	struggling and wanting help or reassurance	involved, informal
5a	regaining confidence	praising the user for getting back on track
5b	unsure and wanting support	patient, praising the user for a difficult success
6	pleased with him/herself, lively	pleased with the user, excited
7	neutral	businesslike, formal

Table 3: Interpretations for the Rules in Table 2

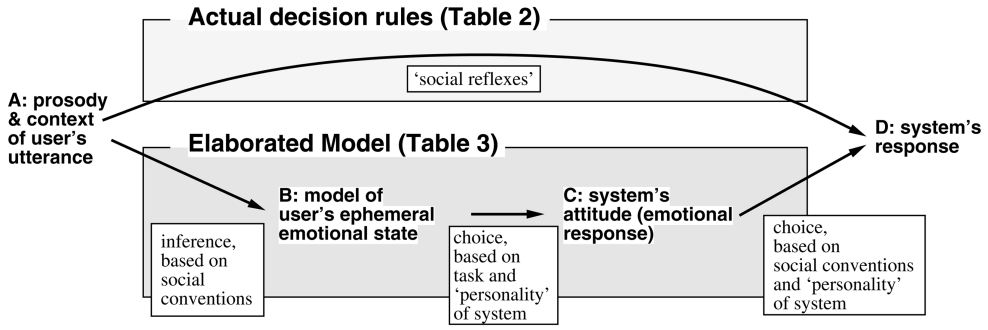

Figure 2. Two Architectures for Responding to Ephemeral Emotions. A, B, C, and D refer to the columns in Tables 2 and 3.

dition would be to respond invariably with *hai*, the most frequent, neutral, and polite acknowledgement. However in the course of developing the rules we found a strong 'variation preference': most people strongly disliked monotonous responses, considering them cold and formal. We accordingly used as a control a version which chose acknowledgements at random, while preserving the frequencies of each acknowledgment in the corpus. This we considered to be 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.

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.

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 listened to the user's guesses and typed `y` or `n`, 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 outputing the acknowledgements, was done by the system. Acknowledgements were pre-recorded samples of a human voice.

A second implementation issue was that of speed. Since ephemeral emotions fade quickly, the window of opportunity for a relevant response is 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 the wizard practiced until he was able to classify most inputs before the user had finished the word, although with occasional errors. Thus the user was never kept waiting. 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 many cases where one of the participants broke the simple guess-confirm routine with a meta-comment, joke or 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.

The most serious problem was that of devising a way to elicit user judgements. In pilot experiments it turned out that the difference between the two versions of the system was hard to get at: simple questionnaires revealed no consistent user preference for one over the other. In part this was to be expected: with users being completely involved in the task of recalling station names, they probably had no attention to spare for judging the quality of the contributions of the system; nor did they have any motivation to do so. Thus the differences between the system acknowledgements probably fell below the level of conscious awareness for most subjects. We believe this is a general problem for systems which operate at near-human levels of performance, or at least without gross infelicities [25]. Over extended use, even over a few minutes, we believe that the cumulative effect of consistently saying just the right thing would give users an overall impression that the rule version was supportive and attentive, and conversely that the cumulative effects of slightly awkward responses in the random version would make users find that version hard to talk to. But 90 seconds was, it appeared, too short for such effects to be felt.

As a work-around, we introduced a second evaluation phase: after interacting with each system the user listened to a recording of his own interaction, while following along on a automatically generated transcript and judging on a seven-point scale the quality of each acknowledgement. After this he filled out a questionnaire asking: which system he would prefer to use if he were to use it for one hour, how he ranked the each two versions on naturalness and various other dimensions, on seven-point scales, and so on. Various cross-tests suggest that this technique of 'evaluation after relistening' is an effective way to amplify weakly-detected user preferences [25].

The subjects were juniors participating in experiments to fulfill a class requirement. Each user interacted with the full system and the random version; the order of presentation was varied. 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 to us, 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. After this we had usable data from 13 subjects.

RESULTS

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 ranked significantly better on the 'naturalness' dimension ($p < 0.05$ 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.

The fact that there were subjects who did not prefer the more responsive system is also interesting. In part this was probably due to chance, to bugs in the rules used in the current system, and to the failure to vary the prosody of each acknowledgement. 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. Maybe personality traits, such as reactions to being monitored and thoughts about personal control [20], are involved here, which of course raises the question of how to detect and adapt to the different interactional styles and preferences of users.

PROSPECTS AND FUTURE WORK

We foresee that the modeling of ephemeral user emotions 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. 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. Many 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 which are stretched or distorted or padded with fillers, as produced by users who speak while they are still thinking.

Second there is the problem of the cues involved in real-time social conventions. A large part of this 3 man-year project was spent identifying the prosodic manifestations of user desires and feelings; this cries out for a systematic analysis and general theory.

Third there is the problem of the ephemeral emotions themselves. Our inventory of these, in Table 3, was arrived at *post hoc*. Again there is a need for a systematic analysis and general theory that can serve to guide the designs of such systems.

Given some basic research in these areas, 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 expressions of ephemeral emotion 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 must be inferred dynamically from context. Moreover the personality that the system is to project, which determines the B-to-C mappings in Figure 2, will also need to vary from system to system.

DISCUSSION

While ephemeral emotions as such have not been identified as a topic in interface research before, they lie at the intersection of two influential dreams: the dream of systems which infer the user's implicit emotional state, and the dream of systems which follow the conventions of social interaction.

The first dream, of exploiting emotions [18, 2], is often proposed 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, there has as yet been no compelling role identified for these, except for entertainment purposes, where the system is intended to be watched more than actually interacted with. We believe that the focus should instead be on the *ephemeral* emotions: because, if the goal is to improve user interfaces, it makes sense to use those emotions which are most related to communication and social interaction.

Dealing with ephemeral emotions moreover avoids some potential problems with inferring the user's emotions: that failure to so accurately may annoy him, but that success may make him feel deprived of control. Regarding failure, for ephemeral emotions failure is not mission-critical: the effects of the inference do not effect subsequent interactions. That is, in case of failure, the system may seem momentarily cold, out-of-synch, non-attentive, foreign, or perhaps robotic, but there is no interference with the content of the exchange. Regarding success, with ephemeral emotions, this need not belittle the user, as the underlying assumption is not that the user is incapable of expressing what he or she wants [15], but that the user is clearly (albeit non-verbally) indicating his or her feelings and needs.

Dealing with ephemeral emotions has another advantage. Since the responses they evoke come so swiftly, users don't

expect anything beyond simple reflex-type responses, and so there is no need for the complex sorts of inference seen in some AI-type user modeling systems. On the other hand, the need to respond swiftly makes implementation harder. This is not just an algorithmic or hardware problem but also an architectural one: perhaps requiring multiple simultaneous threads of control (something not supported by today's standard architectures for dialog management) in order to allow the reactive (shallow, emotion-based, conventional) responses to execute swiftly, and somewhat autonomously from more deliberative, content-based response planning [8].

The second dream is that of building systems which obey social conventions, and especially non-verbal conventions [7]. In the long term, systems which are unable to handle these seem destined to have only limited user acceptance [14]. Recent research in this area 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"). Grunt and Aizula were two systems which exploited prosodic cues in the user's voice to decide when to chime in [21, 30]. Thorisson and Cassell's ([8]) 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. Cassell *et al.*'s ([6]) 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.

Of these systems, however, experimental results have been reported only for Thorisson's system, and even this system was tested without proper controls. As a result, it was unknown, until the results reported here, whether users like these systems because they really implement human social interactional conventions, or whether they like them merely because of the variation preference: the basic human preferences for characters that are more active and exhibit a wider repertoire of actions. Our results show that there is indeed a payoff: using social conventions can result in a system which measurably improves the user experience.

While there are doubts about the general idea of using human-to-human interaction as a model for the design of user interfaces [22], exploitation of the conventions of real-time social interaction is probably immune to these concerns. In particular, 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. Moreover when the behavior variation is subtle and in a separate channel (or 'out-of-band' or 'in a separate modality'), as is generally true for back-channel feedback and non-verbal expressions, this is not interrupting or distracting to users. This is an important advantage of multi-modal systems.

Finally, regarding the general interest in making systems which seem human-like, the handling of ephemeral emotions is an effective but relatively high-cost way to do so. If the goal is merely to create systems which are human-like and "believable", or to 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 [16, 19]. However, 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, is a harder problem entirely.

SUMMARY

We have identified the role of ephemeral emotions in human interaction, argued that they can be important in real-time interactive systems, and verified this by experiment. This result shows that using such social conventions can result in a system which measurably improves the user experience. We learned also 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 understanding of prosodic and non-verbal cues in dialog, and advances in the understanding of ephemeral emotions and their role in human interaction. Nevertheless, we ultimately see this line of work as essential to the development of truly effective interface agents, able to persuade, motivate, charm, and sell.

ACKNOWLEDGMENTS

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

REFERENCES

1. Angles, Jeffrey, Ayumi Nagatomi, & Mieharu Nakayama (2000). Japanese Responses *hai*, *ee*, and *un*: yes, no, and beyond. *Language and Communication*, 20:55–86.

2. Ball, Gene & Jack Breese (2000). Emotion and Personality in a Conversational Agent. In *Embodied Conversational Agents*, pp. 189–219. MIT Press.

3. Bavelas, Janet Beavin, Nichile Chovil, Linda Coates, & Lori Roe (1995). Gestures Specialized for Dialogue. *Personality and Social Psychology Bulletin*, 21:394–405.

4. Brennan, Susan E. & Maurice Williams (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.

5. Brooks, Rodney A. (1991). Intelligence Without Representation. *Artificial Intelligence*, 47:139–159.

6. Cassell, Justine, Tim Bickmore, *et al.* (1999). Embodiment in Conversational Interfaces: Rea. In *CHI '99*, pp. 520–527. ACM Press.

7. Cassell, Justine, Tim Bickmore, *et al.* (2000). Human Conversation as a System Framework. In *Embodied Conversational Agents*, pp. 29–63. MIT Press.

8. Cassell, Justine & Kristinn R. Thorisson (1999). The Power of a Nod and a Glance: Envelope vs. Emotional Feedback in Animated Conversational Agents. *Applied Artificial Intelligence*, 13:519–538.

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

10. Fridlund, Alan J. (1997). The New Ethology of Human Facial Expressions. In J. A. Russell & J. Fernandez Dols, editors, *The Psychology of Facial Expression*, pp. 103–129. Cambridge.

11. Graesser, Arthur C., Katja Wiemer-Hastings, *et al.* (1999). AutoTutor: A Simulation of a Human Tutor. *Journal of Cognitive Systems Research*, 1:35–51.

12. Hatfield, Elaine, John T. Cacioppo, & Richard L. Rapson (1994). *Emotional Contagion*. Cambridge University Press.

13. Johnson, Jeff (2000). *GUI Bloopers: Don'ts and Do's*. Morgan Kaufmann.

14. Johnstone, Anne, Umesh Berry, Tina Nguyen, & Alan Asper (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.

15. Lanier, Jaron (1995). Agents of Alienation. http://www.well.com/user/jaron/agentalien.html.

16. Lester, James C., Sharolyn A. Converse, *et al.* (1997). The Persona Effect: Affective Impact of Animated Pedagogical Agents. In *CHI '97*, pp. 359–366. ACM Press.

17. Oviatt, Sharon (1996). User-Centered Modeling for Spoken Language and Multimodal Interfaces. *IEEE Multimedia*, pp. 26–35.

18. Picard, Rosalind (1997). *Affective Computing*. MIT Press.

19. Reeves, Byron & Clifford Nass (1996). *The Media Equation*. CSLI and Cambridge.

20. Rickenberg, Raoul & Byron Reeves (2000). The Effects of Animated Characters on Anxiety, Task Performance, and Evaluations of User Interfaces. In *CHI '00*, pp. 49–56. ACM Press.

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

22. Shneiderman, Ben (2000). The Limits of Speech Recognition. *Communications of the ACM*, 43:63–65.

23. Tsukahara, Wataru (1998). An Algorithm for Choosing Japanese Acknowledgments Using Prosodic Cues and Context. In *International Conference on Spoken Language Processing*, pp. 691–694.

24. Tsukahara, Wataru (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.

25. Tsukahara, Wataru & Nigel Ward (2000). Evaluating Responsiveness in Spoken Dialog Systems. In *International Conference on Spoken Language Processing*, pp. III: 1097–1100.

26. Walker, M. A., D. J. Litman, C. A. Kamm, & A. Abella (1998). Evaluating Spoken Dialog Agents with PARADISE: Two case studies. *Computer Speech and Language*, 12:317–348.

27. Ward, Nigel (1997). Responsiveness in Dialog and Priorities for Language Research. *Systems and Cybernetics*, 28(6):521–533.

28. Ward, Nigel (2000). The Challenge of Non-lexical Speech Sounds. In *International Conference on Spoken Language Processing*, pp. II: 571–574.

29. Ward, Nigel & Takeshi Kuroda (1999). Requirements for a Socially Aware Free-standing Agent. In *Proceedings of the Second International Symposium on Humanoid Robots*, pp. 108–114.

30. Ward, Nigel & Wataru Tsukahara (1999). A Responsive Dialog System. In Yorick Wilks, editor, *Machine Conversations*, pp. 169–174. Kluwer.

31. Yankelovich, Nichole, Gina-Anne Levow, & Matt Marx (1995). Designing SpeechActs: Issues in Speech User Interfaces. In *CHI '95*, pp. 369–376. ACM Press.