WORKING NOTES, AAAI SPRING SYMPOSIUM ON BELIEVABLE AGENTS
STANFORD UNIVERSITY, CALIFORNIA, AUGUST 13-16, 1994, 86-90

FACE-TO-FACE COMMUNICATION WITH COMPUTER AGENTS

Kristinn R. Thórisson

The Media Laboratory
Perceptual Computing Section
Massachusetts Institute of Technology
20 Ames Street E15-411 Cambridge MA 02139
kris@media.mit.edu

Abstract

While computers are becoming more intelligent, current interaction methods, such as keyboards, mice and windows, still limit human-computer interaction to tool-level manipulation. Bringing a communication paradigm to the computer seems a worthy goal, in particular communication that people use every day to interact with each other. This work begins to attack this issue by examining some of the variables that allow people to conduct fluent and reactive turn-taking and give real-time feedback in everyday face-to-face interactions. A central part of this endeavor is the control of a graphical face that can produce some of the behavior exhibited by people in conversation. For the metaphor to be useful, the behavior of such faces has to be believable.

Keywords: *Agents, face-to-face, real-time, human-computer interaction.*

1. INTRODUCTION

Recently there has been an increased interest in computer interfaces that combine multiple input and output modalities to increase the communication bandwidth with computers [Koons et al. 1993, Bolt & Herranz 1992, Herranz 1992, Mochizuki et al. 1992, Neal & Shapiro 1991, Thorisson et al. 1992, Tyler et al. 1991, Wahlster, 1991, Hauptman 1989, Bolt 1980]. This interest stems from a desire to get away from learned, pre-defined interaction techniques and move towards more flexible, natural ones.

Among the strengths of social communication are its use of *multiple modes* and multiple *information types* and it's inherent *flexibility*. These factors allow people to communicate with each other in many ways, combine complex information in a concise manner and switch dynamically between representational styles. While the first two factors have received attention in the literature for computer interfaces [Koons et al. 1993, Thorisson et al. 1992, Bolt 1984], general flexibility in the input/output sequence has been largely ignored. Yet it may be argued that interaction fluidity on par with human interaction would be extremely beneficial when interacting with machines, since most of the people in the world are "experts" in this style of communication.

One of the problems with constructing flexible multi-modal interfaces is the awkwardness of gesturing, speaking and looking around without having *someone*—or *something*—to address [Britton 1991]. This is partly because current computer interfaces are not endowed with the correct feedback mechanisms—the ones we are used to when talking face-to-face with other people. The obvious solution to this problem is to simulate the social setting to a sufficient extent, including adding an "entity" or embodiment to the interface that the user can address. My work focuses on the issue of creating useful and believable reactive feedback to users in the form of a face that can carry on a real-time face-to-face interaction with them. The final goal of this research is to allow the computer to provide both reactive and reflective behavior for an interface agent, using an interaction style modeled on human dialogue.

2. IMPLICATIONS FOR AUTONOMOUS AGENT DESIGN

Face-to-face interaction has interesting features that set it apart from other interaction methods, the most important one being the number of modes that a person can employ to convey a single thought: facial expressions, various types of gestures, intonation and words, body language, etc. This introduces redundancy into the communication channel that the agent should be able to take advantage of, as well as ambiguity that the agent has to be able to resolve. For believable

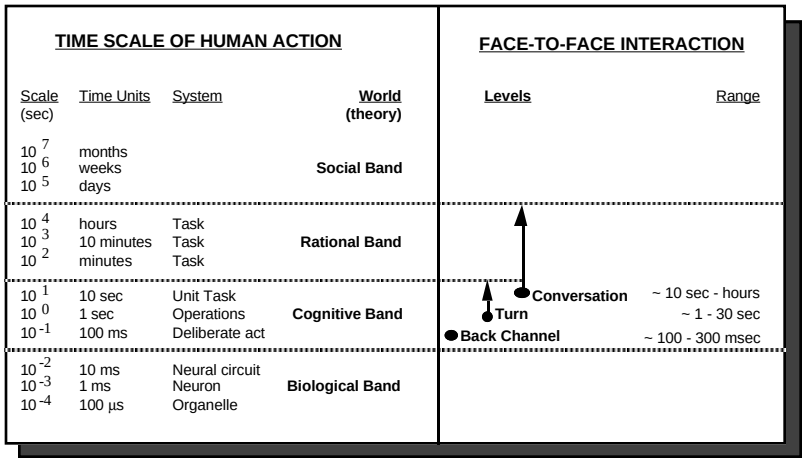

Figure 1. Comparison between the timing in face-to-face interaction and the time scales of human action as classified by Newell [1990].

face-to-face interaction, an agent would probably have to have access to all of the different information channels present in a face-to-face dialogue and represent them in a format useful for generating complementary social behavior.

A less obvious feature of face-to-face communication are the demands that it puts on the timing and management of behaviors. For example, new fixation points are determined on the average of three to four times per second [Card et al. 1983], back channel feedback [Yngve 1970] requires a recognize-act cycle of around 100 ms, and single turns [Whittaker & O'Conaill 1993, Duncan 1972] span somewhat longer intervals. Whole conversations run from a few seconds (quick greetings) to hours. Therefore, to be believable, conversational agents have to be capable of both reactive and reflective behavior. Figure 1 shows how three major parts of dialogue compare to the various time scales of human action identified by Newell [1990].

A third feature of face-to-face interaction important to the design of embodied computer agents is that the *interaction space* be available to the agent's sensory apparatus. This is crucial for generating believable gaze behavior and deictic references. How such environmental "awareness" is achieved depends of course on the implementation; immersive environments—where the user and agent both occupy the same virtual space—are considerably simpler to deal with in this respect than systems where the agents are situated in the real world.

3. J. Jr.: A SOCIAL INTERFACE AGENT

To explore some of the issues relevant to reactive social behavior, a prototype system called J. Jr. was designed [Thorisson 1993]. This system deals specifically with the data and control mechanisms for allowing *real-time social responses* of the agent. I have elsewhere defined *social interface agents* as *agents that are familiar with the conventions of personal interaction* [Thorisson 1993]. This is to distinguish them from other work on agents where the prevalent interaction method is the use of keyboards, mice, windows, and icons [Maes 1993, Kozierok & Maes 1993, Vere 1991, Oren et al. 1990, Chin 1991, Laurel 1990, Crowston & Malone 1988]. To further distinguish social from "animal-based" metaphors, the terms *embodied interfaces* and *personified agents* may be used.

Since social interaction is necessarily multi-modal, the dialogue system in J. Jr. uses data from three input modes: the user's hand gestures, gaze and intonation. Data about gaze and gestures is provided by a human observer in a "Wizard of Oz" manner (a person monitors the user's actions and keys them in[1]); data about intonation in the user's speech is obtained with automatic intonation analysis (see [Pierrehumbert & Hirschberg 1990] for a discussion on intonation). This information is in turn used to automatically control the gaze of J. Jr.'s on-screen face (Figure 2), its back-channel paraverbals, and turn-taking behavior, which consists of asking questions at appropriate points in the dialogue.[2] Examples of the control structures used in this system are given in [Thorisson 1993]. At the risk of oversimplifying input analysis, the system focuses on defining minimum requirements for believable reactive face-to-face behavior.

3.2 An Example Interaction[3]

The current version of the system allows a user to speak in a natural manner to J. Jr. through a microphone. J. Jr. will give back-channel feedback and ask questions at appropriate times in the dialogue. Because the system can "see" the user's hands, it will not interrupt if the user waves her hands around while

[1] Elsewhere we have described and employed automatic methods to gather this data; see [Thorisson et al. 1992] for a discussion of hand-tracking and [Koons & Thorisson 1993] for an eye tracking method designed for estimating line of sight and intersection with real-world objects like computer displays. Eventually this information will be captured by cameras (see e.g. [Essa et al. 1994]).

[2] Since asking questions and saying "mhm, aha" are the exact qualifications for hosting a talk-show, J. Jr. is named after a well known talk-show host. Like any respectable host, J. Jr. asks only questions that are very general and have no relation to what the user is saying.

[3] A VHS (NTSC) video cassette of this interaction session is available from the author.

looking for a word. Non-grammatical pauses ("ahhh...", "uhhh", etc.) will not cause the system to break down since the agent "knows" that the user hasn't finished by the state of the user's speech input. And a pause where the user looks at the ceiling while recalling a word will not cause the system to interrupt, unless that pause becomes excessively long.

The following is a transcript of an interaction between the author and the system. For the interaction to work well the user has to speak in a normal manner—as if addressing a human listener (the intonation filtering will not work otherwise). Since no content analysis is performed the user cannot expect any answers to questions. Not seen in this transcript is the agent's gaze, which is modeled after human interaction [Goodwin 1981] and plays an important role in indicating to the user what the current dialogue state is. The important point to note is the timing of the agent's back channel feedback and turn taking, relative to the user's speech

Key: *Speaker:[beginning, sec:ms] Utterance [ending, sec:ms].*

Three dots (...) mark a pause longer than half a second; commas are pauses that are less than that. The agent's turn taking (and utterance of canned questions) are marked in bold. (Again, note that while the content of all utterances is pre-determined, their timing is not.) Paraverbals are randomly selected every time.

User: [00:000] Hello J. [00:550]
Agent: [01:450] Hi, welcome, nice to see you. [04:100]
U:[09:650] Nice to see you too, you know, I've been ahh [09:650] ...
[10:350] working on you for a long time now and it seems like it's about time that you start behaving. [12:150]
A:[12:950] Yes. [13:400]
U:[14:250] And, ahh, [14:650]
A:[15:150] Yeah [15:600]
U:[15:600] and you seem to be doing that you know. [17:050]
U:[17:600] Oh, by the way, I like that propeller on your hat. It's kind of nice [20:450] ...
U:[21:000] it's kind of
U:[22:000] cute [34:060]
A:[22:050] Yeah, aha. [24:800]
A:[25:700] So what's new? Anything interesting? [28:900]
U:[29:900] Well, you know, just, just the regular stuff [32:850] ...
U:[33:150] playing [33:450]
A:[33:200] Yes? [33:400]
U:[33:550] playing a little racquetball, keeping in shape. [35:800]

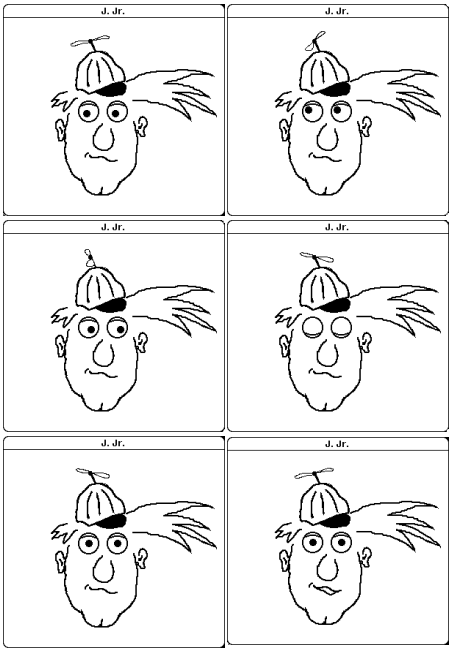

Figure 2. J. Jr.'s on-screen face is capable of looking around, blinking, rotating the hat propeller and opening and closing the mouth in coarse synchronization with synthesized speech.

A:[36:950] Yeah. [39:150]
U:[39:700] Yepp. [40:000] ...
A:[41:450] Well, that's interesting. What else? [44:300]
U:[45:000] I don't know what to tell you, ahh [46:750] (gesticulates) ....
U:[50:450] Well, you seem to be working properly now, that's, that's one good thing. [54:600]
A:[55:050] Yes. [55:750]
U:[56:750] And, ahh, I, I just hope you [58:800]
U:[59:650] gain a little bit more in your intelligence so that ... [1:02:250]
A:[1:02:250] Really? [1:02:750]
U:[1:03:400] Yeah, well, I'd...I'd, I'd like to see you [1:05:950] ...
A:[1:07:050] Aha. [1:07:650]
U:[1:07:450] have a little bit more brains, really [1:09:950]
A:[1:10:100] Yeah? [1:10:350]
U:[1:11:250] and, ahh, [1:11:700]
U:[1:12:400] but otherwise you're fine, [1:13:500]
U:[1:14:02] you know [1:13:950]
A:[1:14:150] Yes. [1:14:350]
U:[1:14:550] I can assure you. [1:15:150]
A:[1:16:350] Yeah. [1:16:600]
A:[1:18:000] Well, thank you for sharing that. [1:19:650]
A:[1:21:250] It was nice talking to you. [1:23:150]
A:[1:23:800] Good-bye. [1:24:400]
U:[1:25:250] Good-bye. [1:25:500]

4. DISCUSSION

The agent described here displays some of the reactive behavior necessary for successful social interaction. First-time users of the system often get the impression that the system makes use of powerful automatic speech recognition and language understanding to produce the observed behavior. This speaks for the relative quality of the turn-taking behavior, giving an informal "context-independent Turing test" for the social behavior of the agent. The system (and the illusion of semi-intelligence) breaks down when users start to speak nonsense to it. This is usually a somewhat disappointing moment and is a clear indication that this type of interface requires somewhat sophisticated speech, gesture and gaze recognition to be even justified. What the required level of sophistication is, however, is not obvious at the present time.

Future work will focus on these issues: adding automatic speech recognition, gesture parsing and line-of-gaze analysis to allow more advanced behavior on part of the agent. By adding speech recognition, automatic speech *generation* also becomes feasible and will allow for a more meaningful interaction. This will necessitate adopting more powerful data handling methods that can deal with interpretation, reaction, and planing in an integrated manner. Among the key issues to this end is the design of a general turn-taking mechanism that can take advantage of redundancy in the various modes and dynamically correct for errors in the communication—and at the same time allow for real-time interaction. Such a mechanism will undoubtedly be a crucial part in coordinating face-to-face interaction and thus creating a truly interactive, embodied agent.

ACKNOWLEDGMENTS

This work has benefited from interactions with Richard Bolt, Pattie Maes, David Koons and Alan Wexelblatt. Joshua Bers and Jennifer Jacoby have my thanks for commenting on these notes. This research is funded by the Advanced Research Projects Agency (ARPA) under Rome Laboratories, contracts F30602-89-C-0022 and FC30602-92-C-0141, and by Thomson-CSF.

REFERENCES

Bolt, R. A. & Herranz, E. (1992). Giving Directions to Computers via Speech, Gesture and Gaze. *Proceedings of UIST '92*.

Bolt, R. A. (1980). "Put-That-There": Voice and Gesture at the Graphics Interface. *Computer Graphics, 14(3)*, 262-70.

Bolt, R. A. (1984). *The Human Interface*. Belmont, CA: Lifetime Learning Publications.

Britton, B. C. J. (1991). *Enhancing Computer-Human Interaction With Animated Facial Expressions*. Master's Thesis, Massachusetts Institute of Technology, Cambridge, Massachusetts.

Card, S., Moran, T. P. & Newell, A. (1983). *The Psychology of Human-Computer Interaction*. Hillsdale, NJ: Lawrence Earlbaum Associates.

Chin, D. N. (1991). Intelligent Interfaces as Agents. In J. W. Sullivan & S. W. Tyler (eds.), *Intelligent User Interfaces*, 177-206. New York, NY: Addison-Wesley Publishing Company.

Crowston, K. & Malone, T. W. (1988). Intelligent Software Agents. *Byte*, Dec., 267-271.

Duncan, S. Jr. (1972). Some Signals and Rules for Taking Speaking Turns in Conversations. *Journal of Personality and Social Psychology, 23(2)*, 283-292.

Essa, I., Darrell, T. & Pentland, A. (1994). Modeling and Interactive Animaiton of Facial Expression using Vision. *M.I.T. Media Laboratory Perceptual Computing Section Technical Report No. 256*.

Goodwin, C. (1981). *Conversational Organization: Interaction Between Speakers and Hearers*. New York, NY: Academic Press.

Hauptman, A. G. (1989). Speech and Gestures for Graphic Image Manipulation. *SIGCHI Proceedings '89*, 241-245. New York: ACM Press.

Herranz, E. Giving Directions to Computers via Speech, Gesture and Gaze. Master's Thesis, Massachusetts Institute of Technology, 1992.

Koons, D. B. & Thorisson, K. R. (1993). Estimating Direction of Gaze in Multi-Modal Context. Paper presented at *3CYBERCONF—The Third International Conference on Cyberspace*, Austin, Texas, May 13-14.

Koons, D. B., Sparrell, C. J. & Thorisson, K. R. (1993). Integrating Simultaneous Input from Speech, Gaze and Hand Gestures. In M. T. Maybury (Ed.), *Intelligent Multi-Media Interfaces*. AAAI/MIT Press.

Kozierok, R. & Maes, P. (1993). A Learning interface Agent for Scheduling Meetings. *SIGCHI International Workshop on Intelligent User Interfaces*, Florida. New York: ACM Press.

Laurel, B. (1990). Interface agents: Metaphors with character. In B. Laurel (ed.) *The Art of Human-Computer Interface Design*, 355-365. Reading, MA: Addison-Wesley Publishing Co.

Maes, P. (1993). Learning Interface Agents. SIGCHI *International Workshop on Intelligent User Interfaces*, Florida. New York: ACM Press.

Mochizuki, K., Takemura, H. & Kishino, F. (1992). Object Manipulation and Layout in a 3-D Virtual Space Using a Combination of Natural Language and Hand Pointing. *SPIE*, Vol. 1828, Sensor Fusion V, 106-113. Bellingham, DC: Society of Photo-Optical Instrumentation Engineers.

Neal, J. G., & Shapiro, S. C. (1991). Intelligent Multi-Media Interface Technology. In J. W. Sullivan & S. W. Tyler (eds.), *Intelligent User*

Interfaces, 11-43. New York: ACM Press, Addison-Wesley Publishing Company.

Newell, A. (1990). *Unified Theories of Cognition*. Cambridge, MA: Harvard University Press.

Oren, T., Salomon, G., Kreitman, K. & Don, A (1990). Guides: Characterizing the Interface. In Laurel, B. (ed.) *The Art of Human Computer Interface Design*, 367-81. Reading, Massachusetts: Addison-Wesley Publishing Company.

Pierrehumbert, J. & Hirschberg, J. (1990). The Meaning of Intonational Contours in the Interpretation of Discourse. Chapter 14 in P. R. Cohen, J. Morgan and M. E. Pollack (eds.), *Intentions in Communication*. Cambridge: MIT Press.

Thorisson, K. R. (1993). Dialogue Control in Social Interface Agents. *InterCHI Adjunct Proceedings*, Amsterdam, April, 139-140.

Thorisson, K. R., Koons, D. B. & Bolt, R. A. (1992). Multi-Modal Natural Dialogue. *SIGCHI Proceedings '92*, April, 1992, 653-4. New York: ACM Press.

Tyler, S. W., Schlossberg, J. L., & Cook, L. K. (1991). CHORIS: An Intelligent Interface Architecture for Multimodal Interaction. In *AAAI '91 Workshop Notes*, 99-106.

Vere, S. A. (1991). Organization of the Basic Agent. *SIGART Bulletin*, Vol. 2, No. 4, 164-168.

Wahlster, W. (1991). User and Discourse Models for Multimodal Communication. In J. W. Sullivan & S. W. Tyler (eds.), *Intelligent User Interfaces*, 45-67. New York: ACM Press, Addison-Wesley Publishing Company.

Whittaker, S. & O'Conaill (1993). An Evaluation of Video Mediated Communication. *InterCHI Adjunct Proceedings*, Amsterdam, April, 73-4.

Yngve, V. H. (1970). On Getting a Word in Edgewise. *Papers from the Sixth Regional Meeting*, 567-78. Chicago Linguistics Society.