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 multimodal 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
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, Kozierek & 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 in1); 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 Interaction3
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

---

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

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]


A: [15:150] Yeah [15:600] and you seem to be doing that you know. [17:050]


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 planning 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


