DIALOGUE CONTROL IN SOCIAL INTERFACE AGENTS

Kristinn R. Thórisson

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

ABSTRACT

Interface agents are computational entities that form a focal point for communication at the interface; *social interface agents* are familiar with the conventions of *personal* interaction. This paper outlines a prototype social interface agent, called J. Jr., that integrates various channels of information about the user to control its real-time behavior in the social setting. Information about the user's gaze and hand gestures is provided by a human observer; data about intonation in the user's speech is obtained with automatic frequency analysis. This data is in turn used to control the gaze of the agent's on-screen face, its back-channel paraverbals, and turn-taking behavior. Results show that by choosing the appropriate variables for dialogue control, a relatively convincing social behavior can be achieved in the agent.

KEYWORDS: Social interface agents, multi-modal dialogue, real-time interaction.

INTRODUCTION

Agents at the computer interface are generally used to represent a given class of procedures and can often autonomously accomplish high-level goals within a limited area of expertise. Social interface agents are knowledgeable about the conventions and rules of personal interaction and allow for social interaction with users, using combinations of speech, gaze, facial gestures and gesticulation. The goal of the current research is to analyze some of the factors controlling social interaction, as a first step towards an intelligent interface that closely mimics human face-to-face interaction. The main vehicles for this study are the idea of an on-screen agent: a visual representation of the computational processes that work collectively to interpret and execute a user's commands, and multi-modal input involving hand gestures, speech and direction of gaze.

Agents

Many people have presented their own visions of interface agents (see e.g. [1]), but until recently they were dealt with more in theory than practice. While reality is still lagging behind the vision, some progress has been made. Two recent systems offer a good contrast to the current work. Oren et al. [3] describe an interface that uses characters to facilitate database access and search. Each agent in this system represents a specific point of view; pre-recorded video segments of real people are the agents' visual representation. A user interacts with the system with a keyboard and a mouse. Vere [5] describes a system that allows for natural language interaction with a simulated autonomous submarine. Commands are given via typed sentences through a keyboard. Both these systems use the idea of agents to facilitate control at the interface. However, both employ an communication style that bears little resemble to interaction in a social encounter.

Multi-Modal Input

One of the strongest arguments for talking to computers is its potential flexibility. However, this flexibility will be partially lost if speech cannot be used along with other natural means of communication. Thus, for high bandwidth interaction we would want to integrate the speech input with automatic hand gesture recognition and analysis of the user's direction of gaze. Recent research in this direction is described in [2] and [4].

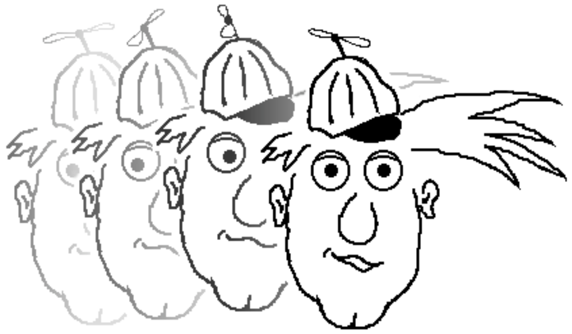

Figure 1. J. Jr.'s visual representation is a face. His behavioral repertoire consists of looking around, looking at the user, speaking, blinking, and rotating the hat propeller. Gaze and synthetic speech are controlled by dialogue states.

Published in: InterCHI Adjunct Proceedings '93, Conference on Human Factors in Computing Systems, Amsterdam, April 24-29, 139-140.

A SOCIAL INTERFACE AGENT

Among the tasks of the social interface designer is to make the computer recognize dynamically when it is appropriate to give back-channel paraverbal feedback (say “mhm”, “aha”, “I see”, nod, etc.), when to respond in speech, and when it should turn away to do what we have asked it. To investigate the issues involved in such real-time dialogue control, a prototype agent, called J. Jr., was designed (Figures 1, 2). J. Jr. is an attempt to extract the features controlling social behavior, specifically paraverbals and turn-taking, without being dependent on an extensive analysis of the dialogue content.

Interacting with J. Jr.

The current implementation of the system, which runs on a Macintosh IIfx computer, allows a user to speak in a natural way to J. Jr. through a microphone. J. Jr. will give back-channel feedback and ask questions at appropriate times in the dialogue.[1] If the user waves his hands around while looking for a word, J. Jr. will not interrupt. If a user pauses with an “ahhh...”, J. Jr. will wait until she is finished. When interacting with the system for the first time, most people get the impression that it requires powerful automatic speech recognition and language understanding. This shows how believable J. Jr.’s back-channel and turn-taking behavior is. Since no such understanding is involved, however, users soon realize (upon talking nonsense to it) that J. Jr. has no idea what the topic is. However, careful selection of the variables controlling interaction results in a relatively convincing dialogue behavior.

The agent’s gaze behavior is a strong indication of whether it seems to be “paying attention” or not. When the eyes wander around aimlessly, most people get the feeling that it cannot “hear” that they are speaking to it. Gaze plays thus a role in indicating “system status”, both for showing dialogue state and the status of the interactive agent.

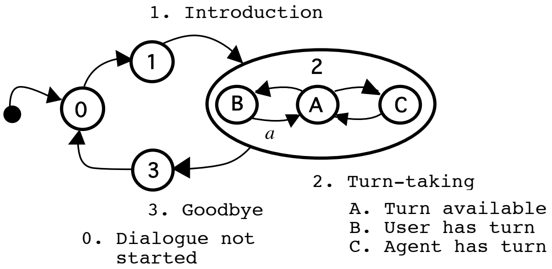

Figure 2. State diagram showing the control structure of J. Jr. Each state has a specific set of actions that the agent is capable of performing, as well as conditions for jumping to the next possible state.

Input and Output Variables

The user’s actions are measured by one time-dependent variable and four Boolean variables: (1) whether the user is *looking at the agent or not* (`look-on/off`), (2) whether the user is *gesturing or not* (`gestures-on/off`), (3) whether the user is *speaking or not* (`speech-on/off`), (4) what the *direction of intonation is* (frequency going *up* or *down*, `pitch-dir = up/down`), and (5) *time* (in msec.) *since the user spoke* (`speech-silence = msec`). A sixth variable, (6) *time since agent looked at user*, is used to control the agent’s gaze behavior. The first two variables are given to the system by a human observer through a keypad, the speech variables are measured automatically using a microphone and a frequency analysis system. All variables are measured and quantified in real-time, parallel to the user’s behavior, thus making real-time response of the agent possible.

Dialogue States

The control structure of J. Jr. can be viewed as a finite state machine augmented with a global clock. Each state (Figure 2) has conditions that variables 1-5 above have to meet for the system to jump to the next state. For example, the rule for jumping between turn-taking states **B** and **A** (arch *a* in Figure 2) is:

```
(OR (AND      gaze-on
              speech-off
              gestures-off
              pitch-dir = down)
         (speech-silence > threshold))
```

where `threshold` is a pre-determined value in milliseconds. If either condition is satisfied, the dialogue will move from state **B** to **A**.

FUTURE EXTENSIONS

The agent described here displays some of the behavior necessary for social interaction. Research on the behavior of the agent should focus on methods for information abstraction to make real-time automatic speech generation possible and allow the agent to respond in a more flexible manner. Future extensions on the input side will include continuous speech recognition, automatic gesticulation analysis and direction-of-gaze estimation.

ACKNOWLEDGEMENTS

The author wishes to thank David Koons, Pattie Maes and Richard Bolt. This research was in part funded by DARPA, contract F30602-89-C-0022.

REFERENCES

1. *The Art of Human Computer Interface Design* (1990). Laurel, B. (ed.). Reading, Massachusetts: Addison-Wesley Publishing Company.

[1] Since asking questions and saying “mhm, aha” are exactly the necessary qualifications for hosting a talk-show, J. Jr. is named after a well known American talk-show host. Like all respectable hosts, J. Jr. asks only questions that are very general and have no relation to what the user says.

Published in: InterCHI Adjunct Proceedings ’93, Conference on Human Factors in Computing Systems, Amsterdam, April 24-29, 139-140.

2. Koons, D. B., Sparrell, C. J. & Thórisson, K. R. (in press). Integrating Simultaneous Input from Speech, Gaze and Hand Gestures. To be published in M. Maybury (ed.), *Intelligent Multi-Media Interfaces*. AAAI Press.
3. Oren, T., Salomon, G. & Kreitman, K. (1990). Guides: Characterizing the Interface. In Laurel, B. (ed.) *The Art of Human Computer Interface Design*, 367-81. Reading, Massachusetts: Addison-Wesley Publishing Company.
4. Thórisson, K. R., Koons, D. B. & Bolt, R. A. (1992). Multi-Modal Natural Dialogue. *SIGCHI Proceedings '92*, April, 653-4. New York: ACM Press.
5. Vere, S. A. (1991). Organization of the Basic Agent. *SIGART Bulletin*, Vol. 2, No. 4, 164-168.

Published in: InterCHI Adjunct Proceedings '93, Conference on Human Factors in Computing Systems, Amsterdam, April 24-29, 139-140.