J.Jr.: A Study in Reactivity

6.

The J.Jr. system [Thórisson 1992] was a pilot system designed to explore the idea of reactive multimodal behavior in an interface agent. This system served as a precursor to the development of Ymir and highlights important problems in multimodal dialogue, which will be addressed at the end of the chapter.

6.1 System Description

In the J.Jr. system, dialogue control in is based on an FSM (finite state machine), augmented with a global clock. It 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 according to a pre-determined scheme); data about intonation in the user's speech is obtained with automatic frequency analysis (Figure 6-1). This information is in turn used to control the gaze of J. Jr.'s on-screen face (Figure 6-2), its back-channel paraverbals, and turn-taking behavior, which consists of asking questions at appropriate points in the dialogue.[1]

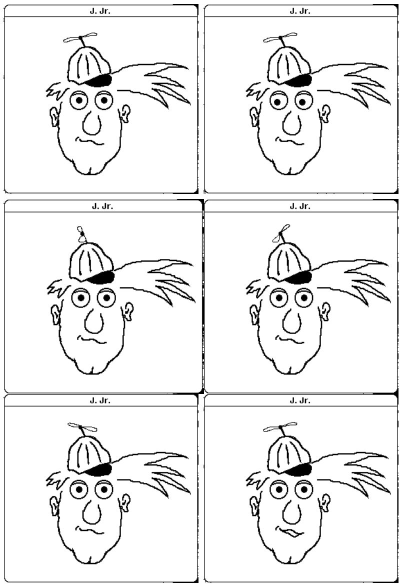

FIGURE 6-2. J.Jr.'s face is capable of looking around, blinking, rotating the hat propeller and opening and closing the mouth inrough synchronization with synthesized speech.

6.1.1 Input: Gestures, Gaze & Intonation

In the J.Jr. system gestures and gaze are quantified into Boolean variables; if the line of gaze intersects the on-screen agent face, the variable GAZE-ON? is set to TRUE , else it is FALSE. If the user moves his or her

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

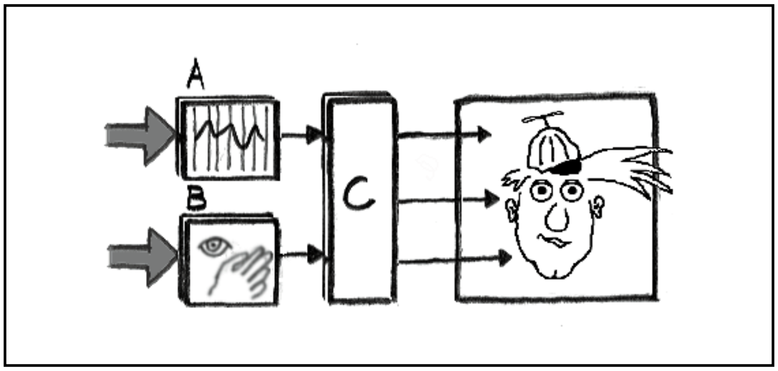

FIGURE 6-1. System structure. of J.Jr. The user's speech is automatically processed for intonational constituents and pauses (A). Information about gaze and gestures are monitored and input through a keypad by a human observer (B). The dialogue system (C) controls the cartoon character's speech, gaze and hat propeller

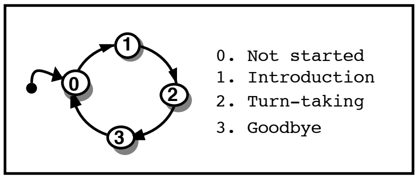

FIGURE 6-3. State diagram showing the control structure of the social encounter in J. Jr. Each state has a specific set of actions that the agent is capable of performing, as well as conditions (see text) for jumping to the next possible state.

hands in a way that obviously relates to the dialogue (i.e. excluding "self adjusters"—fixing of the hair, scratching, etc. [Rimé & Schiaratura 1991]), the variable GESTURES-ON? gets a TRUE value, else it is FALSE. This relatively sophisticated analysis of gesture is possible by using a human observer to code the user's behavior in real-time.

Pierrehumbert & Hirschberg's [1990] work strongly indicates that intonational features are important indicators about the intentional and structural features of discourse. Utterances contain intonational phrases made up of combinations of high and low pitches. Phrases can be divided up into sub-phrases, or intermediate phrases, which contain relatively small variations in pitch. The intonational phrase as a whole ends with either an increased high or low. In the J.Jr. system a simple filter is used to detect whether the speaker's pitch is rising or falling. Other speech variables used were SPEECH-ON? which is given a TRUE if the user is speaking, otherwise FALSE; and Silence, which contains the time in milliseconds since the user spoke. A third variable, PITCH-DOWN?, is set to TRUE if the intonation is falling, otherwise, if the intonation is rising or stays constant, it takes on a FALSE value.

The variable SILENCE contains the time in milliseconds since the user spoke. This turns out to be a very important element to time the actions of the agent.

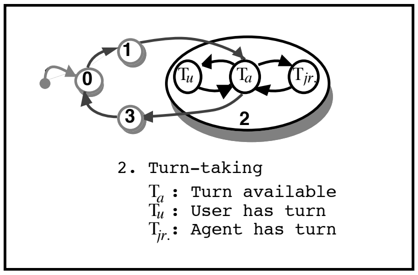

FIGURE 6-4. State diagram showing the three dialogue states, Tu, Ta, and Tjr, embedded within "encounter" state 2. In the original implementation the agent always asks the user a question in state Tjr.

6.1.2 Output: Speech, Turn Taking, Back Channel, Gaze

The agent's gaze and back-channel behavior is controlled with two variables: BACK-CHANNEL-ALLOW? and LOOK-AT-USER. The variable BACK-CHANNEL-ALLOW? is set to TRUE only when the user has turn and is used to control when the agent gives back-channel feedback [Yngve 1970]. It is also used to prevent multiple paraverbals in a row, by setting it to FALSE immediately after a paraverbal has been given and waiting for the user to continue before resetting it to TRUE .

As discussed before ("Gaze" on page 44), results of research on gaze behavior in multi-modal interaction [Goodwin 1981] shows that the eyes play an important role in turn-taking; a speaker looks away at the beginning of his or her utterance, but as the utterance approaches termination gazes back to the recipient. The variable LOOK-AT-USER controls the gaze of the agent and is set to TRUE at appropriate points in the dialogue. (Looking at the user is accomplished by having the face look straight out of the screen.) If this variable is FALSE, the agent looks around at random.

6.1.3 Dialogue States

The dialogue control mechanism is a finite state machine augmented with a global clock. There are four "general" states for the dialogue encounter, and three for the turn-taking or dialogue itself (Figure 6-3 & Figure 6-4; the encounter states are numbered from 0 to 3).

Encounter state 2 is divided into three sub-states, or turn taking states, shown in Figure 6-4. These are marked Tu, Ta, and Tjr, for "user has turn," "turn available" and "agent has turn," respectively. Transitions between the states requires certain conditions to be true, determined by the values of the input variables.

6.1.4 State Transition Rules

In the following discussion a state change is denoted Change-State [a → b] or simply [a → b], where a is the prior state and b is the new state. The interesting states to look are the turn-taking states and how the agent achieves back channel feedback (state 2 in Figure 6-4). The initial sub-state is Ta. To make the transition [Ta → Tu], the simple condition **R1** (Figure 6-5).

The constant DIALOG-UNITS is set to 100 ms. This is the smallest unit of time measurement in the system; all other thresholds are multiples of this value. To go back to Ta we look for the conditions shown in **R2**, where (* 5 DIALOG-UNITS) is a multiplication of the constant Dialog-Units by five. For the agent to take the turn ([Ta → Tjr]) we wait for situation **R3** to arise.

```
R1: User Takes Turn

IF  (OR
        speech-on?
        gestures-on?)
THEN
    (Change-State [Ta → Tu]))

R2: User Gives Turn

IF  (OR
        (AND
          look-on?
          pitch-down?
          (not speech-on?)
          (not gestures-on?))
        (> Silence
          (* 5 Dialog-Units)))
THEN
    (Change-State [Tu → Ta])
    (Look-at-User ← TRUE)

R3: Agent Takes Turn

IF  (OR
        (AND
            (> Silence
                (* 2 Dialog-Units))
            look-on?))
        (> Silence
            (* 6 Dialog-Units))
THEN
    (Change-State [Ta → Tjr])
    (Look-at-User ← FALSE)
    (Turn-Propeller)
    (Ask-Question [Next-Q])
    (Change-State [Tjr → Ta])

R4: Agent Gives Back Channel

IF  (AND
        allow-back-channel?
        (not gestures-on?)
        (> Silence
            (* 1.1 Dialog-Units))
THEN
    (Give-Back-Channel)

(allow-back-channel? ← FALSE)
```

FIGURE 6-5. Pseudo code control algorithms for J.Jr.'s turn taking and back channel feedback behaviors.

The agent will look away and rotate the hat propeller as a clue to indicate that he is taking the turn. Since the agent cannot use body language to indicate dialogue states, these turn out to be fairly useful cues for the user. The dialogue goes back to state Ta, [Tjr → Ta], immediately after the agent has finished the question. The function Ask-Question takes an argument, NEXT-Q, which contains the question to be vocalized by the speech synthesizer. This is read from a canned script of questions.

6.1.5 Back Channel Feedback

The back-channel mechanism is the only behavior to make use of the smallest unit of time measurement in the system, DIALOG-UNITS, which is set to 100 msec. In state Tu the rule **R4** (Figure 6-5) will produce back-channel feedback from the agent: The variable ALLOW-BACK-CHANNEL? is set to TRUE when entering state Tu. It is set to FALSE immediately after the back-channel feedback has been given, and back to when the user has started speaking again if the state is still Tu. The multiplier for DIALOG-UNITS in this case will undoubtedly vary depending on the "pace" of the dialogue, but judging from research on humans (see "Back-Channel Feedback" on page 40), is unlikely to need to be less than 1.0.

6.2 Discussion

First-time users 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 and back channel, giving an informal "context-independent Turing test" for the dialogue behavior of the agent. A real interaction scenario with J. Jr. is described in Figure 6-6. While this system shows that accurate timing, intonation and crude gesture/gaze analysis can provide a sufficient mixture to take turns correctly, it also points to the problems of creating extensive systems that integrate reactive abilities with higher-level competence.

6.3 The Problem with J.Jr.

The system (and the illusion of semi-intelligence) breaks down when users start to speak nonsense to it—usually a somewhat disappointing moment for users, but not at all unexpected to the designer. I refer to the problems typified in this system as [1] the sensing problem, [2] the lack of behaviors problem, [3] the reactive-reflective integration problem, and [4] the expansion problem.

6.3.1 The Sensing Problem

Using a human observer to classify the kinds of gestures that the user does totally bypasses the problem of automatic gesture classification. Even though morphemic features of body motions are relatively gross, compared to intonation for example, they still may be difficult to analyze automatically because of the phenomenon of morphemic substitutability ("Morphological and Functional Substitutability" on page 76). One of the inherent problems lies in selecting the correct time-scale to analyze a person's behavior on. The importance of determining simple features like whether the user is addressing the computer agent or another person, whether a vocalization is a filler or an actual utterance that contains semantic information, cannot be stressed enough. These are the features that make system behavior robust.

6.3.2 The Lack of Behaviors Problem

A human conversant has a wealth of behaviors to choose from. On any occasion, these are chosen based on various features of the dialogue, and they are chosen in real-time. J.Jr. provides only a simple mapping between a state and its behaviors, but more importantly has no way to select or compose alternative multimodal acts if it did (this should perhaps be called the arbitration problem).

6.3.3 The Reactive-Reflective Integration Problem

How would we integrate natural language understanding into the J.Jr. system? If we want to integrate the content of utterances with intonation analysis and body language, we have to deal with complications like delayed production of results, backtracking time of occurrence of events and guaranteeing response (see "Computational Characteristics of Psychosocial Dialogue Skills" on page 65). How are we to integrate inofrmation content with real-time process control? When should a user utterance like "huh?" spin off a process that tries to re-plan a previous utterance? These are questions of internal and external process control, and they covary closely with the methods we employ for extracting information from the multimodal input stream. An outline of a solution to these problems will be provided in the next chapter.

6.3.4 The Expansion Problem

By using a finite state machine (FSM) as the basic mechanism of dialogue tracking, a serious limitation is set to the amount and ease of expansion. This means that building complex characters, with hundreds of behaviors (from blinking to planning many kinds of utterances), will be extremely difficult and time comsuming. FSMs are good for tracking states, and clearly we want to keep track of states in any dialogue system. But for anything else in a dialogue system, perception, action control, multimodal integration, FSMs are not the right kind of mechanism, firstly because these processes are hard to describe in terms of states and their transitions, and secondly because the complexity of multimodal dialogue requires an incremental approach to behavior building, and FSMs don't lend themselves easily to such an approach. A possible solution to low-level (reactive) behavior would seem to be something like Brooks' [1986] subsumption architecture, but no clear mechanism exists in that approach to deal with higher-level analysis and output generation.

Kris: [00:000] Hello J. [00:550]
J.Jr.: [01:450] Hi, welcome, nice to see you. [04:100]
K:[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]
J:[12:950] Yes. [13:400]
K:[14:250] And, ahh, [14:650]
J:[15:150] Yeah [15:600]
K:[15:600] and you seem to be doing that you know. [17:050]
K:[17:600] Oh, by the way, I like that propeller on your hat. It's kind of nice [20:450] ...
K:[21:000] it's kind of
K:[22:000] cute [34:060]
J:[22:050] Yeah, aha. [24:800]
J:[25:700] So what's new? Jnything interesting? [28:900]
K:[29:900] Well, you know, just, just the regular stuff [32:850] ...
K:[33:150] playing [33:450]
J:[33:200] Yes? [33:400]
K:[33:550] playing a little racquetball, keeping in shape. [35:800]
J:[36:950] Yeah. [39:150]
K:[39:700] Yepp. [40:000] ...
J:[41:450] Well, that's interesting. What else? [44:300]
K:[45:000] I don't know what to tell you, ahh [46:750] (Butterworth gesture) ...
K:[50:450] Well, you seem to be working properly now, that's, that's one good thing. [54:600]
J:[55:050] Yes. [55:750]
K:[56:750] And, ahh, I, I just hope you [58:800]
K:[59:650] gain a little bit more in your intelligence so that ... [1:02:250]
J:[1:02:250] Really? [1:02:750]
K:[1:03:400] Yeah, well, I'd...I'd, I'd like to see you [1:05:950] ...
J:[1:07:050] Aha. [1:07:650]
K:[1:07:450] have a little bit more brains, really [1:09:950]
J:[1:10:100] Yeah? [1:10:350]
K:[1:11:250] and, ahh, [1:11:700]
K:[1:12:400] but otherwise you're fine, [1:13:500]
K:[1:14:02] you know [1:13:950]
J:[1:14:150] Yes. [1:14:350]
K:[1:14:550] I can assure you. [1:15:150]
J:[1:16:350] Yeah. [1:16:600]
J:[1:18:000] Well, thank you for sharing that. [1:19:650]
J:[1:21:250] It was nice talking to you. [1:23:150]
J:[1:23:800] Good-bye. [1:24:400]
K:[1:25:250] Good-bye. [1:25:500]

FIGURE 6-6. Sample dialogue between the author and the J.Jr. system. Notice that while the content of all utterances is pre-determined, their timing is not. Paraverbals are randomly selected every time.

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.