Requirements for a Socially Aware Free-standing Agent

Nigel Ward and Takeshi Kuroda[1]
University of Tokyo, School of Engineering
nigel@sanpo.t.u-tokyo.ac.jp, kuro@sanpo.t.u-tokyo.ac.jp

Abstract

As robots venture forth into human spaces, they will have to interact with busy strangers on the move. In order to do so they will need to incorporate models of human-human social interaction. This paper proposes a testbed for developing and prototyping such models: a two-dimensional agent system placed in a lobby where it can interact with passers-by. This paper reports our preliminary analysis of the social interactional abilities such a system must have. It further discusses some of the gesture, posture, and position signals that the system must detect and respond to in real-time, based on analysis of videotaped records of people interacting with a mock-up of the system.

Robots and Social Interaction

If you want a robot to hit ping-pong balls, you need to know how ping-pong balls move. If you want a robot able to interact with people, you need to know what people do. This is true especially for robots with humanoid shapes or abilities, which implicitly invite people to treat these robots as they treat other people. Unfortunately, the science of human-human interaction is incomplete in some of the areas most relevant to the engineering of communicative systems.

This has not yet held back robotics research, for two reasons. First, many users are able to adapt to the style of interaction required by the computer. For example, Polly (Horswill 1993) required a person to shake a leg if he wanted to continue the interaction. But, looking forward to the day when robots proliferate, out from laboratories filled with technologically savvy people and into places where they interact with the populace at large, they will have to deal with people who are less willing or less able to respond in idiosyncratic ways. Thus the need for humanoid robots with human-like styles of interaction.

Second, if the inventory of actions of the robot is limited, interaction is inevitably simple — all that user input is effectively used for is to select among the possible actions the robot can perform. In such cases, the user may as well choose from a menu. However, if a robot can do more it has to acquire more information from the user. In particular, a robot that engages in cooperative action with a user will has to be able to handle communication that manages the status of the ongoing activity. This is clear in human-human interaction, for example two people moving a table give signals to each other throughout, indicating the degree of exertion they are making, their stability, their next planned action, and their perceptions of how close they are to the goal. Even in purely verbal interaction this is clear: two people who are discussing something indicate continually how interested or uninterested they are with the current topic, how long before they expect the discussion to continue, whether they want to take charge of the discussion or be more passive, how well they are understanding the other person, and so on.

In human-human communication, this is typically done without recourse to explicit actions; for example, it is rare for people to say "I need to shift my left foot so don't push right now" or "I think this conversation is getting off the topic" (except in cases where the other person has failed to pick up on the more subtle, normal signals). Robots should similarly be 'socially aware', able to pick up such subtle signals.

We are interested in building systems that can communicate in these ways. One important pre-

[1] We thank Keishin Tanigawa, Kuan Liang Loh, and Takashi Kobayashi for their ideas and help. This work was supported in part by the Inamori Foundation, by the International Communications Foundation, and by the Okawa Foundation for Information and Telecommunications.

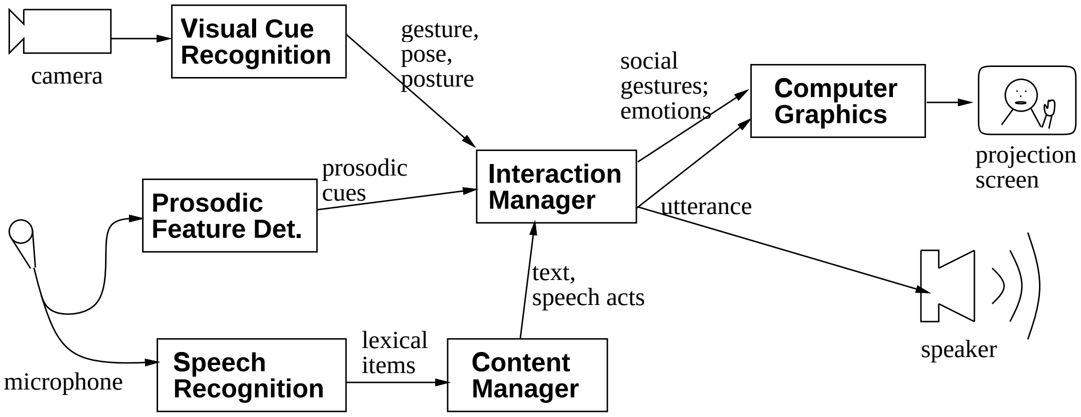

Figure 1: System Components

requisite is requirements analysis, to discover what social interactional factors a robot must be sensitive to in order to communicate naturally, and this is the first goal of this paper. Given such an inventory of factors, the next step is to determine how to acquire the relevant information. The second goal of this paper is thus to uncover the visual cues to these factors, including gesture, position, and posture.

Our Task

Rather than studying all of social interaction and all visual cues, we focus attention on those aspects most relevant to humanoid robots, based on the following reasoning:

Humanoid robots will, in all likelihood, encounter many people in the course of their activities. Thus a substantial part of the robot's work will involve the initiation and termination phases of interaction, that is, the process of greetings and farewells. These aspects of interaction have, unfortunately, not been well enough studied as yet; much previous work in human-computer action is focused on what happens after the interaction begins, and assumes that the user is sitting down and facing the system, or is even wired up to the system (Thorisson 1996). We assume rather that people are standing and free to move throughout the interaction.

Humanoid robots may, moreover, need to initiate interactions, not just wait passively for a person with a specific need to arrive, and wait for him formulate a query. Moreover they should not require a human operator to be present to give advice on how to use the system, nor require a placard to tell people what to expect from the robot. Rather they should be 'free-standing', able to call out to people, and to lead conversations.

Humanoid robots will need to deal with people who are busy, and so cannot assume that the user is paying attention exclusively to the system. In particular, they must allow for the possibility that the user is simultaneously doing other things, thinking about other things, and even talking and gesturing with other people. While simple models of man-machine interaction generally assume that every signal which the user produces is produced in response to a system action, and is directed to the system, humanoid robots will not have this luxury.

As a testbed for these abilities, we have chosen to develop an agent system for the lobby of a university building to interact with the people going in and out. Since our concern is with the forms of the interactions more than their content, our system will not be tailored to perform any specific useful ability. Rather it will specialize in initiating, engaging in, and gracefully terminating short interactions. The content of the interactions may include 1. small talk about the weather, 2. jokes, 3. a description of our department, delivered at a pace adaptive to the user's response (Iwase & Ward 1998), 4. a simple memory game in which the system has the user name the countries of the European Union, or the stations of the Yamate-loop line, while providing feedback and hints (Tsukahara 1998), and 5. having the user explain something complex, while encouraging him to continue by means of generating back-channel feedback ("uh-huh") at appropriate times (Ward & Tsukahara 1999). We chose these topics because they allow the possibility of interesting interactions despite simple content, and can be implemented without the need for accurate speech

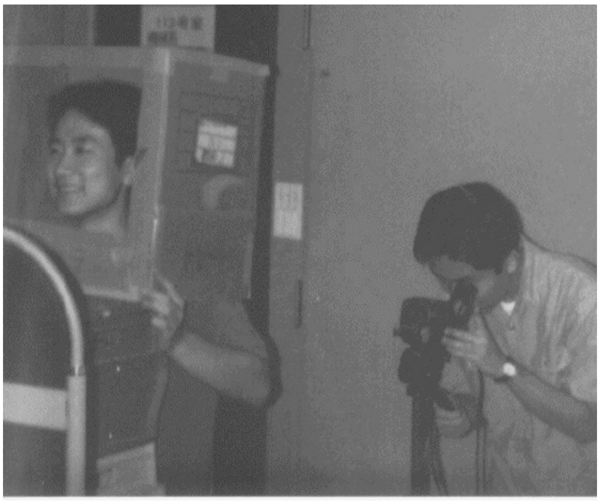

Figure 2: First System Mock-up and Camera-man

recognition, which is difficult in a noisy lobby.

The tentative hardware and software configuration is shown in Figure 1. Note that there will be no actual robot, just a larger-than-life figure projected on a wall screen.

Thus our conception is similar to that behind the August system (Gustafson *et al.* 1999). August was placed in the public area of an exposition hall, emulating an opinionated character who also had a small store of factual knowledge about his creators, local restaurant locations, etc. Its output modalities were synthesized speech and an animated face, capable of lip movement and of head and eyebrow movements synchronized with the stresses in the output sentences. The input devices were a video camera, apparently used for detecting user presence, a microphone for speech recognition, and a push-to-talk button.

Another similar system is the Digital Smart Kiosk (Christian & Avery 1998). This was placed at the entrance to a cyber-cafe and provided information about the cafe. Its output modalities were speech (synthesized or pre-recorded), a web browser display and an animated face (used for smiling and other simple gestures, and for speaking text with synchronized lip movements). The input devices were a video camera, used for detecting and tracking people in order to decide when to greet and where to look, and a touch-screen, used for user button selections.

Our aim is to build a system that extracts more visual information about the user, such as his posture and gestures, and uses it to select, adjust, start or halt system activities in real-time, thus providing a more responsive and natural user experience.

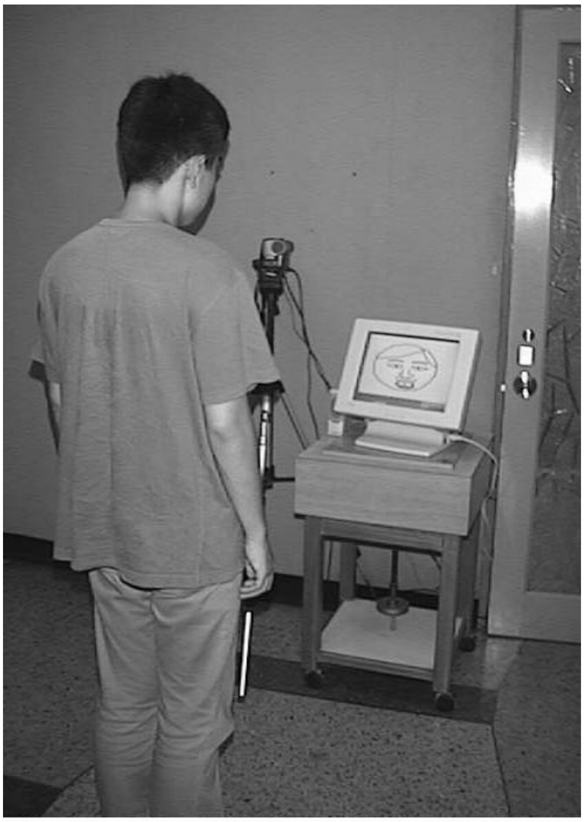

Figure 3: User with Second System Mock-up

Methods

In order to develop our specification for the system we have videotaped people interacting with mock-ups of the system described above. In our first session one of us stood in the lobby of our building with a box on his head, suggesting a computer display (Figure 2). As people crossed the lobby the man in the box would call out greetings and, if the person was willing, engage in a minute or two of banter. Our second and third sessions used sketches of a face presented as a simple slideshow on a display, with a microphone and speaker transmitting audio to and from the operator, who was hidden behind a door (Figure 3). In the third session the operator roughly followed a pre-determined script. So far we have tabulated a total of 143 gestures from 8 users in the second session, and casually examined interactions with 16 other users in the other sessions. The frequency of gestures seemed to drop in sessions where people were talking to a computer screen rather than to a real person.

We then analyzed the videotapes. Our analysis was biased in several ways. First, we ignored gestures that were content-related, which accentuated, illustrated, or substituted for words. These were in any case rare in our data. Rather we focused on gestures with a direct social function. Second, we assumed that gestures with props (cigarettes, handkerchiefs) were equivalent to the same gestures without props. Third, we did not look for complex conjunctive gestures (such as "person places hand is on hip AND turns obliquely"), rather we assumed that each individual gesture would have an independent significance. Fourth, we tended to ascribe meaning to gestures in terms of general dimensions of meaning, attitude, or intention.

Basic Observations

Facial gestures were not common in our data, occurring less often than movements of the head, hands, body, etc.

Posture and position also seemed to be significant. These differ from ordinary gestures in that they last over many seconds, but they seem to bear similar functions, so below we treat then together.

Many of the gestures we observed were highly ambiguous, but nevertheless seemed to have social significance, including gestures such as hair smoothing, nose rubbing, cheek scratching and stretching, which may seem at first glance to be purely self-directed actions.

Greeting and Parting

In the above simulation, the operator had to make many choices. We therefore focused analysis on points where the operator chose to behave in one way rather than another, and tried to relate that choice to gestures from the other person. Sometimes this choice was fairly directly determined other, being almost reactive (Ward 1997). This section and the next list some the operator's actions and the gestures of the other which seemed to cue them, in bracketed italics.

g1 When/whether to greet someone at a distance, with "chotto" (hey), "konnichiwa" (hello), etc. [*close, walking slower than 1.4 m/s*[2], *facing system*]

g2 When/whether to invite them to talk, with "hima desu ka" (do you have a minute?), "chotto hanashimasenka" (can we talk for a minute?), etc. [*close, stopped or approaching, facing system*]

g3 When to open the conversation after the person approaches, with "konnichiwa" (hello). [*very close, bowing*]

g4 When to wrap up the current topic or introduce a new one. *[moving slightly away, shifting weight to other foot, swaying*]

g5 When to start closing the conversation, with "arigato gozaimashita" (thank you). [*starts moving away or turning away*]

g6 When to close the conversation "mata, ne" (see you), etc. *[moving away, waving bye-bye*]

Turn-taking

The operator also had to handle turn-taking.

t1 While talking, at every instant, what to do next:

- keep talking [*nodding*]
- laugh [*laughter (note that laughing is variously indicated by voice, by face, by shoulder shaking, by turning to the side and/or moving the hand in front of the mouth (women), and by bending back at the waist)*]
- stop talking [*sharp, forceful gestures*]
- change the topic

t2 While listening, at every instant, what to do next: (These choices seem to depend largely on the prosody of the other person's utterances, with gesture playing a lesser role.)

- listen silently
- back-channel, with "un" (uh-huh), etc.
- laugh
- conclude he wants a short response and take a short turn, saying "hai" (yes), etc. [*a person rotating his head down as he ends a question expects the system to agree with them*]
- conclude he is finished and take a turn, perhaps introducing a new topic

User States

Sometimes gestures from the user serve as direct signals, telling the system exactly how to behave, that is, what choice to make, as seen above. More often, however, the relation between user

[2] determined by K. L. Loh

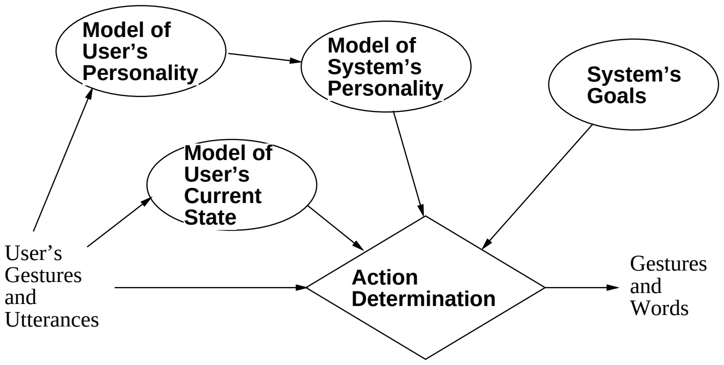
Model of User's Personality
Model of System's Personality
System's Goals
Model of User's Current State
User's Gestures and Utterances
Action Determination
Gestures and Words

Figure 4: Structure of Decision-Making (in the Interaction Manager of Figure 1)

gesture and system action is fairly indirect. It seems that most user gestures indicate something about his internal state, or about his opinion of the state of the interaction. Knowledge of this state can allow the system to make an appropriate choice at the next decision point that comes up. On later passes through the videotapes we therefore focused on aspects of the users's state which the system needed to know, and classified gestures accordingly:

s1 attentive [*crossing arms, hand on hip*]
¬s1 distracted (by another person etc.) [*turning to the side, looking around*]

s2 tense *[rocking body from side-to-side or back-and-forth, touching face or head with hand; a general high level of small motion (fidgeting)*]
¬s2 relaxed

s3 passive and willing to let the system lead *[laughing, nodding while talking, touching head and face*]
¬s3 aggressive and wanting to lead *[talking without gesturing, getting closer while talking, hand gestures away from the body, chest level hand gestures, smoothing hair on top of head*]

s4 wanting to talk [*hands out with palms up, at belly level*]
¬s4 wanting to listen [*nodding*]

s5 interested in the current topic (surprise being the extreme case) *[facing towards system, coming closer, nodding agreement; touching face quickly (in case of surprise)*]
¬s5 uninterested [*turns away slightly or significantly, keeping still, raising hand to look at watch*]

s6 charmed [*laughing*]
¬s6 offended [*straightening posture*]

s7 suspicious [*lowering head to peer at or behind the experimental equipment*]

s8 following, understanding [*nodding*]
¬s8 not hearing, not following [*bending forward, nods and other responses delayed*]
¬s8′ needing time to think *[speaking slowly; looking to one side, putting hand on hip*]

s9 feeling distant, interacting formally
¬s9 feeling close, friendly [*large movements, fast movements*]

s10 dominant
¬s10 submissive, apologetic, polite [*touching back of head or neck*]

Note that terms like 'dominant' and 'submissive' are here intended to indicate transient states; for example, the videotape included a back-hair-smoothing-gesture while making a suggestion; we interpreted this as indicating not a general submissiveness, but as marking that specific suggestion as being a tentative, polite one, not a forceful or commanding one.

The Role of Personality

Since people prefer to interact with systems which manifest a personality similar to their own (Reeves & Nass 1996), a system should attempt to infer a user's personality. Clearly some aspects of user state, discussed in the previous section, may persist over time, in which case they can be considered aspects of the user's personality. Thus, to the extent that a user consistently uses gestures of one kind of another, a system can probably draw inferences about his personality.

A system should then adopt an appropriate personality of its own. This determines many things, including how the system responds to user gestures in the future, as suggested by Figure 4. The rest of this section mentions a few of the things that will need to be affected by personality.

One thing will depend on the system's personality is its choice of wording. For example, a serious system might say things like "we're doing a survey on what facial features an agent system should have", and a more lighthearted one might say "here's my new face, what do you think?". An aggressive system might ask direct personal questions like "where are you going?", while a less aggressive system might offer up only generic comments like "the lobby has been busy today". Vocabulary and grammar must also be chosen to reflect the formality or informality of the chosen personality.

Personality will also determine conversation structure, for example, whether or not the system asks "do you have time to talk to me?" before initiating a topic.

Personality will also need to affect low-level properties, such as: the rate of the system's output in terms of words per second, its tone of voice in terms of spectral qualities and prosodic patterns, its delivery in terms of length of pause between utterances, its quickness in terms of whether it allows long silences or always speaks whenever the other person is not speaking, and its level of redundancy in terms of clarity of explanations.

Previous Research on Gestures

Gestures in social interaction have been a favorite topic of research for psychologists, sociologists, and anthropologists, and are frequently addressed in the robotics, AI, and human-computer interface communities. Previous work can be classified into five main lines of research:

The first line of work focuses on explicit signals (emblems), such as waving bye-bye, raising the arms to signal hooray, shaking the head to mean "no", and so on. These signals typically occur in isolation, as explicit messages. Robots have been built that can recognize and respond to such signals; but the simplistic techniques that suffice here may not be adequate for the more subtle signals that are more common, and more indispensable to coordination control.

The second line of work is that on referential gestures and other gestures that occur during narrative; that is, the gestures that people use to illustrate their words. While there exist systems that can generate such signals (Cassell *et al.* 1994), understanding them is a long-term challenge.

The third line of work focuses on facial gestures encoding emotions, and many computer systems have been built to detect or synthesize such gestures. However emotional displays are rare in normal daily interactions, at least in the cultures we are interested in.

A fourth topic of research is touching as a social gesture (as opposed to touching for position guidance). Although this has has received comparatively little attention, its role for indicating simple approval or disapproval has been exploited in many systems, from the Furby on up.

A fifth line of research addresses gestures which cue turn-taking. Most of this work, unfortunately, focuses on the role of gaze (Goodwin 1981; Duncan & Fiske 1985; Herlofsky 1985; Sakamoto *et al.* 1996), which can be detected reliably only with special hardware (Thorisson 1996).

Our problem, that of reading gestures so that a robot engaged in social interaction can select and adjust its behavior in real time, thus falls outside the purview of all of these research traditions.

Prospects

Above we have inventoried some of the aspects of the user's state that a social agent must pay attention to. Although this list was based on a small amount of data, and is in clear need of refinement, we believe that it has some generality, in part because it coincides with the inventory of aspects revealed by choice of dialog-particles (Ward 1998) such as "so", "okay", "uh-huh" and "um". That is, the same basic information about the social interaction seems to be the object of both gestural and vocal signals.

We have listed some of the gestures that seem to be informative for a social agent. The analysis is tentative, in part because of individual differences in gesturing style, and because of the intrinsic ambiguity of many gestures. To draw firm conclusions it will be necessary to measure correlations between meaning and gesture across larger data sets.

We have assumed that most movements have communicative significance, but have not shown this. The significance of body and hand movements would be easy to measure, by measuring whether people in conversation find videoconferencing more valuable than telephone conversations even if the facial information is obscured.

There may also be more direct ways to measure the communicative significance of hair-smoothing etc.

Although some of the gestures we list are clearly specific to Japanese culture, most probably have the same value for Americans, and perhaps in other cultures. This too needs further study.

This paper has not addressed the question of how a system can recognize gestures. We are optimistic, for two reasons. First, as suggested above, we hypothesize that facial gestures, which are the most difficult to recognize, may be of secondary importance. One reason that this hypothesis is plausible is that people can interact quite well even in cases where they do not directly look at their partner's face, as when their main focus of attention is some other task or when they are working side-by-side. Second, most of the gestures noted involve rather large movements in large regions. Thus it may suffice to use simple algorithms for detecting coarse events, such as swift motion around the lower face (touching the face with the hands), slower larger motions in the head area (nodding), slow repeated body motions (rocking back and forth), breaks in symmetry (turning the head), and so on. This too is plausible; it seems that in daily life people can interact successfully using only peripheral vision, while attending primarily to something else.

Another open question is how important it is to be responsive to gestures. Indeed, it is conceivable that interactive social gestures, although present and meaningful, can be ignored with little or no effect on system performance. After we build our system it will be easy to measure this the value of perceiving and responding to gestures.

This paper has been focused on the question of recognizing gestures produced by the other person. There remains reverse task, of producing gestures to convey information to the other person. Here our analysis is incomplete; in particular, we have not explored the possibility of reflex-like gesture production. For example, it may be that when the other person nods the system must also nod, regardless of its internal state or its inference of the other person's internal state. We also plan to examine this. In any case, however, the basic gesture-meaning correspondences outlined above should be useful for gesture production too.

In the course of pursuing these questions, we plan to build a fully automatic system that can interact naturally with busy moving strangers. This will open the way to smart kiosks and also serve as a prototype for socially competent mobile humanoid robots.

References

Cassell, Justine, Matthew Stone, *et al.* (1994). Modeling the Interaction between Speech and Gesture. In *Program of the Sixteenth Annual Conference of the Cognitive Science Society*, pp. 153–158.

Christian, Andrew D. & Brian L. Avery (1998). Digital Smart Kiosk Project. In *CHI-98*, pp. 155–162.

Duncan, Jr., Starkey & Donald W. Fiske (1985). The Turn System. In Starkey Duncan, Jr. & Donald W. Fiske, editors, *Interaction Structure and Strategy*, pp. 43–64. Cambridge University Press.

Goodwin, Charles (1981). *Conversational Organization: Interaction between speakers and hearers.* Academic Press.

Gustafson, Joakim, Nikolaj Lindberg, & Magnus Lundeberg (1999). The August Spoken Dialogue System. In *Proceedings of Eurospeech 1999.*

Herlofsky, William J. (1985). Gaze as a Regulator in Japanese Conversations. *Papers in Japanese Linguistics*, 10:16–33.

Horswill, Ian (1993). Polly: A Vision-Based Artificial Agent. In *AAAI-93.*

Iwase, Tatsuya & Nigel Ward (1998). Pacing Spoken Directions to Suit the Listener. In *International Conference on Spoken Language Processing*, pp. 1203–1206.

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

Sakamoto, Kenji, Haruo Hinode, & Fumio Togawa (1996). Multimodal interaction model controlling responses based on nonverbal information. In *14th Spoken Language and Discourse Workshop Notes (SIG-SLUD-14)*, pp. 9–15. Japan Society for Artificial Intelligence. in Japanese.

Thorisson, Kristinn R. (1996). *Communicative Humanoids: A Computational Model of Psychosocial Dialogue Skills.* PhD thesis, Massachusetts Institute of Technology, Media Laboratory.

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

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

Ward, Nigel (1998). The Relationship between Sound and Meaning in Japanese Back-channel Grunts. In *Proceedings of the 4th Annual Meeting of the (Japanese) Association for Natural Language Processing*, pp. 464–467.

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