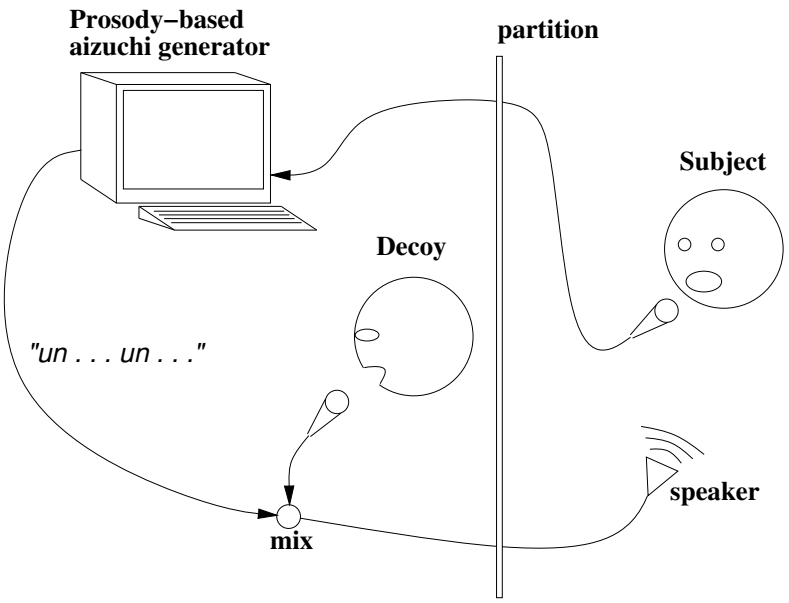

Figure 4: Experiment Set-up

	hate?
because I hate to do that.	really.
	mm
I'm angry that's why.	I'll ask out
	you've decided!
some other girl, I've decided.	<inhalation>,
	well.
I'm 80 percent sure <sigh>.	

Figure 3: Free Translation of the Conversation in Figure 2

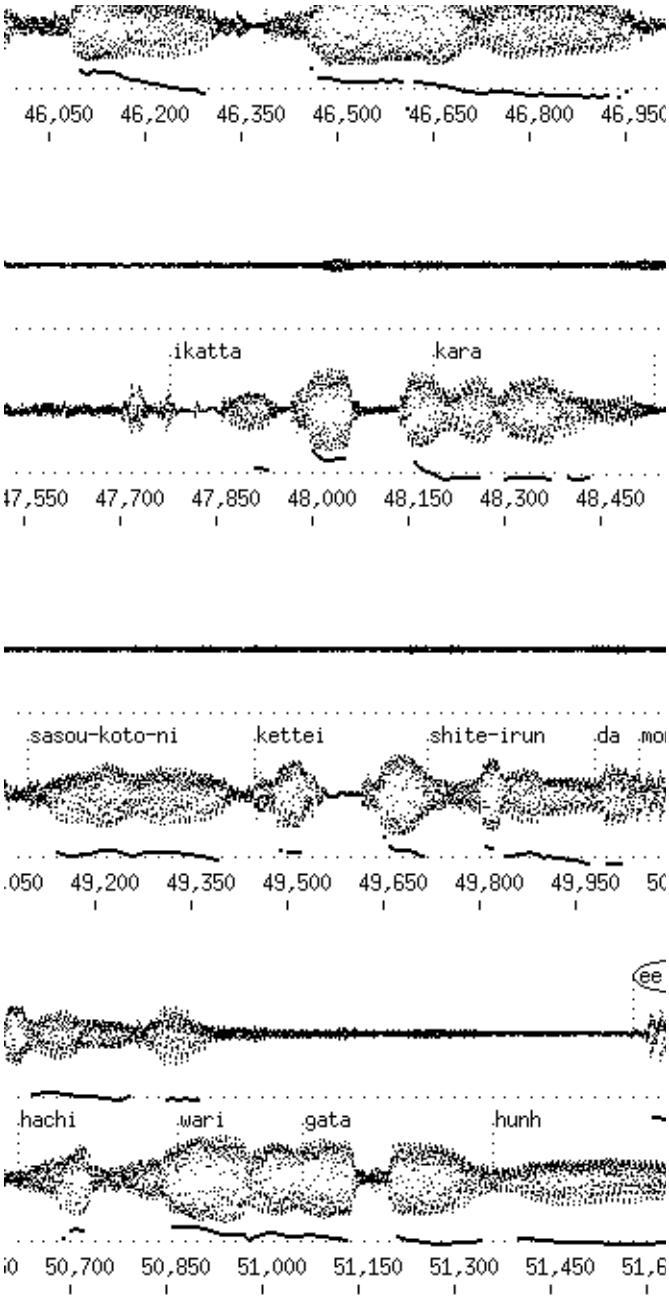

Figure 2: Conversation Fragment. Each of the four strips includes two rows, the top being conversant JH and the bottom being conversant KI. Each row includes: transcription, signal, pitch, and the 30th percentile pitch level (horizontal dotted line). Narrow ovals indicate actual back-channel feedback, and the wide ovals indicate predicted back-channel feedback; note that 3 out of 4 cases were correctly predicted. The context is that tomorrow is KI's birthday but his girlfriend doesn't seem to be planning anything; in the previous utterance he rejected the suggestion that he forget his pride and just ask her to arrange something. A translation appears as Figure 3.

Upon detection of

a region of pitch less than the 30th-percentile pitch level and (C1)

continuing for at least 150ms, (C2)

coming after at least 700ms of speech, (C3)

you should produce back-channel feedback 300ms later, (C4)

providing you have not done so within the preceding 1.2 seconds. (C5)

Figure 1: Back-channel feedback prediction rule for Japanese.

Schegloff, Emanuel A. (1982). Discourse as an Interactional Achievement: Some Uses of "Uh huh" and Other Things that Come Between Sentences. In D. Tannen, editor, *Analyzing Discourse: Text and Talk*, pp. 71–93. Georgetown University Press.

Schmandt, Chris (1994). *Computers and Communication*. Van Nostrand Reinhold.

Ward, Nigel (1996a). Artificial Intelligence and Other Approaches to Speech Understanding: A case study in methodology. In *10th Annual Conference of the Japan Society for Artificial Intelligence*, pp. 403–406. also available by ftp from ftp.sanpo.t.u-tokyo.ac.jp/pub/nigel/papers/jsai96.ps.Z.

Ward, Nigel (1996b). Using Prosodic Clues to Decide When to Produce Back-channel Utterances. In *1996 International Conference on Spoken Language Processing*, pp. 1728–1731.

Yngve, Victor (1970). On Getting a Word in Edgewise. In *Papers from the Sixth Regional Meeting of the Chicago Linguistic Society*, pp. 567–577.

6 In Conclusion

It is sometimes said that Brooks' approach to AI, although fine for low-level behaviors, is irrelevant to higher cognitive processes, such as language. But language use is not just about conveying meaning: for one thing responsiveness is also important, and here Brooks' approach seems very applicable. This is fortunate, since responsiveness and other low-level language behaviors are fundamental to natural human communication, and will be essential for the development of generally usable dialog systems.

References

AAAI (1996). Embodied Cognition and Action: Papers from the 1996 AAAI Fall Symposium. Technical Report FS-96-02, AAAI Press, Menlo Park, California.

Brooks, Rodney A. (1991a). Intelligence Without Reason. In *Proceedings of the Twelfth International Joint Conference on Artificial Intelligence*, pp. 569–595.

Brooks, Rodney A. (1991b). Intelligence Without Representation. *Artificial Intelligence*, 47:139–159.

Brooks, Rodney A. & Lynn Andrea Stein (1993). Building Brains for Bodies. A. I. Memo 1439, Massachusetts Institute of Technology, Artificial Intelligence Laboratory.

Chapman, David (1991). *Vision, Instruction, and Action.* MIT Press.

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.

Evans, Christopher S. & Peter Marler (1995). Language and Animal Communication. In Roitblat & Meyer, editors, *Comparative Approaches to Cognitive Science.* MIT Press.

Frantz, G. A. (1994). Toys That Talk: Two Case Studies. In *Applied Speech Technology*, pp. 487–500. CRC Press.

Maynard, Senko K. (1989). *Japanese Conversation.* Ablex.

Nagao, Katashi & Akikazu Takeuchi (1994). Social Interaction: Multimodal Conversation with Social Agents. In *Proceedings of the Twelfth National Conference on Artificial Intelligence*, pp. 22–28.

Regier, Terry (1996). *The Human Semantic Potential: Spatial Language and Constrained Connectionism.* MIT Press.

- on the other hand, being an appreciative listener, that is, producing appropriate back-channel feedback, for any of the above stereotypical story types, and
- participating in ritual conversational openings ("Hi Joe, how you doing? ..."), and
- engaging in small talk about the weather, etc.

In each case we will rely on prosody as much as possible, do word recognition only as needed, and rely heavily on context to limit the perplexity (the number of likely input types at any point). We also plan to use keyword spotting in most cases, and hope that we can exploit prosody to spot the key (stressed) words from each input.

One problem we face is how to interface low-level behaviors with a word recognizer. The difficulty is that today's best (HMM-based) recognition packages are not designed with the needs of real-time responsiveness in mind. We may have to adapt these models or use other algorithms for recognition.

Fourth, we are interested in the questions of how to understand multi-modal input and how to produce multi-modal output. Both of these are commonly addressed today with classical AI architectures, using lots of representations and lots of symbol manipulation, and striving for full integration. A more workable approach is clearly needed.

We plan to study the possibilities here by building an automated buddy-for-crossword-puzzle-solving, built around a touch screen. Although the system will pretend to help the user do the puzzle, this will be just a pretext for interacting naturally, and hopefully enjoyably, with the user.

Fifth, in addition to building systems, we want to look more at human behavior. One direction is to use back-channel feedback as a source of data on how long various language-related processes take. For example, we may may able to determine that to paraphrase and echo back some new information takes on average more time than to produce an emotional or sympathetic response, or conversely. This may provide a new technique for psycholinguistic research, permitting the study of processing times in real-world, integrated language processing, not just in laboratory studies of comprehension in isolation and production in isolation.

Another direction is to explore whether human language processing should be modeled with a subsumption architecture. Here we plan to look at cases of apparent dissociation among production processes, including stuttering, speech errors, false starts and pauses (which, although typically filtered out and not perceived by listeners, are relatively common), and dissociation among understanding processes, as manifested in inappropriate or sub-optimal responses.

targeted (Brooks & Stein 1993). This section mentions a few things high on the agenda for my students and me, organized mostly around systems we would like to build.

First, we plan to look at back-channel feedback in languages other than Japanese. So far we have come to believe that essentially the same link between low-pitch and back-channel feedback is present in English, for most speakers. This is seen in both corpus data and in live experiments, although we have not yet analyzed enough data to report quantitative results.

Second, we plan to extend our system, retaining the basic strategy of responding passively and without regard to topic. We will start by looking at cases where the current rule fails. This should lead us to the other factors involved in back-channel feedback and turn-taking, and we will then add behaviors to handle these. Iterating the process of failure analysis and system extension will lead us slowly up towards higher level behaviors.

Some specific factors that we will need to isolate in the short term are those which distinguish between back-channel feedback cues, turn-end signals and question signals. Here again we expect to find prosodic cues and to link these fairly directly to appropriate responses. We may also be able to easily measure the speaker's speech rate and degree of emotional involvement; these will also affect the frequency and choice of word and pronunciation for back-channel feedback. We will also look for prosodic clues that indicate what type of feedback the speaker wants: evaluation, opinion, agreement, sympathy, expression of interest, indication of the degree of novelty of the information expressed, etc.

Our goals for the system will be, as before, keeping the conversation going and fooling subjects into thinking they are talking with a human.

Third, we plan to work on single-topic dialogs. Here the system will be able to take a more active role, and will be able to produce targeted responses, not just Eliza-style non-committal responses. Example tasks are:

- participating in naming and memory games (I-Spy-With-My-Little-Eye, or "Can you name all 50 of the United States? I'll keep track and give you hints if you get stuck ..."),
- participating in synchronized chants ("Eenie-Meenie-Mynie-Moe", etc.),
- telling stereotypical stories (telling a joke, giving directions, telling a story about a shopping experience, etc.) and using back-channel feedback from the human listener to determine pacing etc., somewhat in the spirit of Schmandt (1994),

T5 By analogy to the appearance of intelligence emerging from interaction with the world, we see that $\longrightarrow$ the appearance of understanding can arise from interaction with humans.

T6 By analogy to the idea that, for robots, surviving is a more basic goal than performing useful work, it seems that $\longrightarrow$ for dialog, being responsive and keeping the interaction going are more basic goals than understanding and replying with useful information.

T7 By analogy to the attention paid to ethology, we see that $\longrightarrow$ building speech systems requires attention to the facts of human social behavior.

T8 By analogy to the use of subsumption to integrate behaviors in robotics $\longrightarrow$ it may be possible to use subsumption to integrate various basic language behaviors (some discussed below).

The research direction suggested by these tenets is quite different from that taken by mainstream work on spoken dialog systems. Today it is common to: require humans interlocutors to speak clearly, insert pauses, use only certain words, speak as if reading aloud, etc. (vs. T2), focus on the problem of "speech understanding" and pursue it with a more-or-less classical AI methodology (Ward 1996a; Brooks 1991a; Brooks 1991b) (vs. T3), postulate many representations of structure and meaning to mediate between perception and action (vs. T4), focus on complex, derived abilities such as taking dictation and answering questions about data in tables (vs. T6), and prototype the possibilities for computer-human interaction without first studying the nature of human-human interaction (vs. T5 and T7).

More positively, there seems to be a good match between Brooks' approach and a focus on the use of language in social interaction.

However one component of the approach, physical embodiment (T1), seems to be of secondary importance here. This is not really surprising, since it is of course common for people to interact verbally without seeing or touching each other. This is not to say that embodiment is completely irrelevant — even back-channel feedback is clearly related to gaze, nods, etc. (Duncan & Fiske 1985) — but in the short term, we can build language users which are disembodied and still reap most of the benefits of the Brooks approach.

5 Research Directions

So it seems that Brooks' methodology is a good way to study language in social interaction, at least in one case, and probably in others. Indeed, related phenomena, including turn-taking phenomena, prosodically cued phenomena, and grunt-type responses, have already been

3. The decoy switched on the system.

4. After switch-on the decoy's utterances and the system's outputs, mixed together, produced one side of the conversation. The system contributed back-channel feedback (in the form of *un* in a falling pitch) and the decoy contributed follow up questions etc. if needed to keep the conversation going.

I've done the experiment several dozen times informally, as an exhibition at a symposium, at open houses, and also with whoever happens to visit the lab. In every case the system gives a strong impression of responding like a human. It is robust to noise from fans, other conversations, etc., and even works over the telephone. It works both with close friends of the decoy and with total strangers. It does better than a version which produces feedback at random; the latter sometimes sounds bad, and occasionally seems to disrupt and discourage the speaker.

This is the perception of spectators. However naive subjects fail to notice any difference between random and proper back-channel feedback: in neither case do they detect that their conversation partner is partially automated. This suggests that people do not always pay attention to back-channel feedback. To what extent this holds true in general will require further experimentation, perhaps with longer runs, perhaps with different intructions to encourage participants to notice feedback, or perhaps with more sensitive metrics of feedback quality, such as "in which case did you feel the other conversant was being more attentive/friendly/understanding".

4 Responsiveness in Dialog and Brooks-style AI

To summarize how this system relates to Brooks' approach:

T2 By analogy to demonstrations of robots that interact with real, unsimplified worlds, here we see that $\longrightarrow$ it is possible to build dialog systems that interact with real, unconstrained humans.

T3 By analogy to the idea that building a complete world model is impossible and unnecessary, we see that $\longrightarrow$ accurate recognition of words and phonemes is not necessary as a prerequisite to being responsive.

T4 By analogy to motion behaviors based directly on touch sensors, without use of a representation of the world, we see that $\longrightarrow$ direct linking of prosodic perception and verbal action can be effective.

dicted 68% (54/79) of the occurrences of back-channel feedback produced by JH in response to KI, with an accuracy of 76% (54 correct predictions / 71 total predictions), where a prediction is counted as correct if its initiation point is within 500ms of the initiation point of a human-produced back-channel feedback.

Performance was especially good for this conversation, probably because the speakers were very involved in the conversation and because at most points the roles of speaker and listener were clearly distinct. However the rule is also mostly valid for other pairs and other conversation types: running the rule on all 18 conversations in the corpus gave a coverage of 50% (461/914) and an accuracy of 33% (461/1394). For comparison, a rule which used a random predictor instead of C1 and C2 of Figure 1, while obeying clauses C3, C4 and C5, achieved a coverage of 21% (191/914) at an accuracy of 24% (191/807) (best of three runs).

A preliminary study of the errors showed that: A third (roughly 343 of 933) of the incorrect predictions were at places where it seemed that back-channel feedback could have been appropriate. This probably is due in part to differences in back-channel feedback style among speakers; probably no single rule could accurately predict the behavior of all speakers. Another 107 of the incorrect predictions were cases where the rule predicted back-channel feedback at points where the human instead chose to take the floor and produce a full turn. Another 106 or so could be attributed to cases where both speakers were talking at once.

The remaining incorrect predictions, and the missing predictions, are probably due in large part to the fact that the rule is not yet optimized — in particular the clauses of the rule Figure 1 are currently all hard constraints — and to the lack of consideration of meaning.

3 Conversations with Humans

In order to find out how well the above rule would perform in live conversation, I built a real-time version of the system.

[Figure 4 about here.]

To get people to interact naturally with the system it was necessary to fool them into thinking they were interacting with a person. I therefore used a human decoy to jump-start the conversation, and a partition so that the subject couldn't see when it was the system that was responding (Figure 4). The typical procedure was:

1. The subject was told "please have a conversation with this person, and we'll record it so we can add it to our corpus".

2. The decoy steered the conversation to a suitable topic.

2 The Case of Back-Channel Feedback in Japanese

Responsiveness in human dialog is pervasive yet seldom noticed. An interesting example of responsiveness is the *mm*, *uh huh*, and other grunts and words that people produce while listening (Yngve 1970) — that is, "back-channel feedback". Back-channel feedback had been taken as prototypical of social interaction more generally (Schegloff 1982).

Back-channel feedback seems a very basic, low-level part of language in several ways. First, it doesn't require much effort: you can produce feedback without paying attention, and do so well enough avoid detection, for at least 5 or 10 seconds, as when reading a newspaper while someone is talking to you. Second, it doesn't require recognition or reasoning: a one-year-old with no lexical knowledge can sometimes produce back-channel feedback well enough to give the impression of keeping up his end of a conversation. Third, it is not uniquely human: some animals engage in an analogous activity, "dueting" (Evans & Marler 1995). Fourth, it does not take much time: back-channel feedback is produced extremely quickly, often near the minimum human reaction time. So it seems that back-channel feedback is a fundamental aspect of conversation.

I have primarily worked with Japanese, since back-channel feedback in Japanese occurs approximately twice as frequently as in English (Maynard 1989).

I have found that, in Japanese, back-channel feedback is cued by the speaker with a prosodic cue, namely a region of low pitch. (Precursors to this discovery and other related research are discussed in (Ward 1996b).)

The existence of this rule is not something that native speakers are consciously aware of. Most think that back-channel feedback is cued by new information in the speaker's utterance or by the speaker's pauses. However, after the low-pitch cue is pointed out, it is apparent to the unaided ear, when listening to either recorded or live conversations.

[Figure 1 about here.]

Refining this observation with detailed analysis has led to the specific rule in Figure 1. This is the best rule found so far, although improvements are certainly possible. Implementation details appear in (Ward 1996b).

[Figure 2 about here.]

[Figure 3 about here.]

To evaluate the performance of the rule, I compared its predictions to human performance. Figure 2 shows a fragment of one of the conversations in our corpus (Ward 1996b), showing the correspondence between actual human-produced back-channel feedback and rule-predicted back-channel feedback. Over the whole of this 5 minute conversation, the rule correctly pre-

T2 AI must be situated; the robots must operate in the real world, not an artificial simplified one.

T3 AI requires adequate perception, not perfect perception.

T4 AI can be reactive, that is, based on simple behaviors that are fairly direct links between perception and action, not mediated by representation and reasoning.

T5 AI can emerge from simple behaviors, thanks to interaction with the outside world.

T6 AI research should be bottom-up and grounded, starting from simple, essential, low-level behaviors.

T7 AI should be informed by studies of human or animal behavior, not by introspection.

T8 AI can sometimes be achieved with subsumption architectures.

These tenets have inspired various approaches to the study of language. For example:

- T1 suggests that, since much of language use is metaphoric, and many metaphors are based on the human body, an artificial language-user should be embodied in a humanoid robot (Brooks & Stein 1993).
- T1 also suggests that, to exploit the fact that humans are used to interacting with speakers who have bodies, speech systems should be deployed in such bodies (Frantz 1994), or at least with rendered faces (Nagao & Takeuchi 1994).
- T6 suggests that a good place to start studying language is with the most grounded aspects of language, especially those linked closely to perception, such as language for describing simple shapes and their spatial relations, as used in speech-input enhanced drawing tools and in learning models (Regier 1996).
- T8 suggests that another good place to start studying language is its effects on lower-level motion behaviors (Chapman 1991).
- T2, T3, T4, T5, T6, T7 and T8 suggest that language in social interaction, not just language for talking about things, is also interesting to study.

This last topic, language in social interaction, is the focus of this paper. It reports on some analysis and an experiment in this area, explains its relevance to the above tenets, and discusses directions for research.

to appear in Systems and Cybernetics, Special Issue on Embodied Artificial Intelligence and Artificial Life

Responsiveness in Dialog and Priorities for Language Research

Nigel Ward[1]

University of Tokyo

Mech-Info Engineering, 7-3-1 Hongo, Bunkyo-ku, Tokyo Japan 113

fax: +81-3-3815-8356

nigel@sanpo.t.u-tokyo.ac.jp

http://www.sanpo.t.u-tokyo.ac.jp/~nigel/

Abstract

In human language there seems to be a direct link between perception and action for back-channel feedback, that is, the grunts and other acknowledgements produced by a listener in a conversation. In particular, in Japanese, detection of a low pitch region in the speaker's utterance often cues the listener to produce back-channel feedback. This fact has been exploited in a system that can interact with unsuspecting speakers and appear to converse, in real time. This finding suggests that Brooks' (1991a,1991b) methodology for the study and synthesis of physical behavior is relevant also for social behavior. If so, then real-time responsiveness, rather than better phoneme recognition, is a top priority for the construction of spoken language systems and for the scientific study of language.

1 Introduction

Brooks' (1991a,1991b) program for robotics and artificial intelligence (AI) research has inspired and guided much interesting work. Some of the key tenets and tendencies of this methodology, sometimes called the "Embodied AI" approach (AAAI 1996), are:

T1 AI should be embodied in robots.

[1]Acknowledgements: Wataru Tsukahara helped with the analysis and commented on this paper. Keikichi Hirose provided the pitch tracker. Joji Habu helped with experiment design. Tens of students contributed as data sources, as labelers, as decoys, and as subjects. The Sound Technology Promotion Foundation provided funding.