

Reactive Responsiveness in Dialog

Nigel Ward¹

University of Tokyo

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

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

Abstract

This paper discusses implications of Embodied AI for the study of language, reports on a system that can appear to converse, and sketches out a research program for language as used in social interaction.

1. Embodied Artificial Intelligence and Language

Some of the key tenets and tendencies of the Embodied Artificial Intelligence (AI) approach (taken mostly from (Brooks 1991)) are:

- T1** AI should be embodied in robots.
- 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 suggest 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, since humans interact best with speakers who have human-like bodies, speech systems should have such bodies (Frantz 1994), or at least rendered faces (Nagao & Takeuchi 1994).

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

T6 suggests that a good place to start studying language is with the most grounded aspects of language, such as the relation between language and perception, for example when used for describing simple shapes and their spatial relations, as in speech-input enhanced drawing tools and learning models (Regier 1996).

T8 suggests that another good place to start studying language is its effects on lower-level behaviors (Chapman 1991).

T2, T3, T4, T5, T6, T7, T8 suggest that language in social interaction, not just language for talking about things, is also interesting to study. This is the topic of this paper.

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* and *uh huh* that people produce while listening (Yngve 1970; Schegloff 1982). This “back-channel feedback” seems basic in several ways. 1. 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. 2. Even a one-year-old with no lexical knowledge can be responsive enough to give the impression of keeping up his end of the conversation. 3. Some animals engage in an analogous activity, “dueting” (Evans & Marler 1995). 4. Back-channel feedback is produced extremely quickly, often near the minimum human reaction time.

Japanese is a convenient language in which to study back-channel feedback since it quite prevalent, occurring approximately twice as frequently as in English (Maynard 1989). Japanese even has a non-technical term for back-channel feedback, “aizuchi”, which I will use below.

I have found that in Japanese back-channel feedback is cued by the speaker with a prosodic cue, namely a region of low pitch (Ward 1996b), as detailed in Figure 1.

The existence of this rule is not something that native speakers are consciously aware of. Most think that aizuchis are cued by the presence of new information 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.

Upon detection of	(C1)
a region of pitch less than the 30th-percentile pitch level and	(C2)
continuing for at least 150ms,	(C3)
coming after at least 700ms of speech,	(C4)
you should say <i>un</i> 300ms later,	(C5)
providing you have not done so within the preceding 1.2 seconds.	

Figure 1: Back-channel feedback prediction rule for Japanese. This is only the best rule found so far, improvements are certainly possible. Implementation details are given in (Ward 1996b).

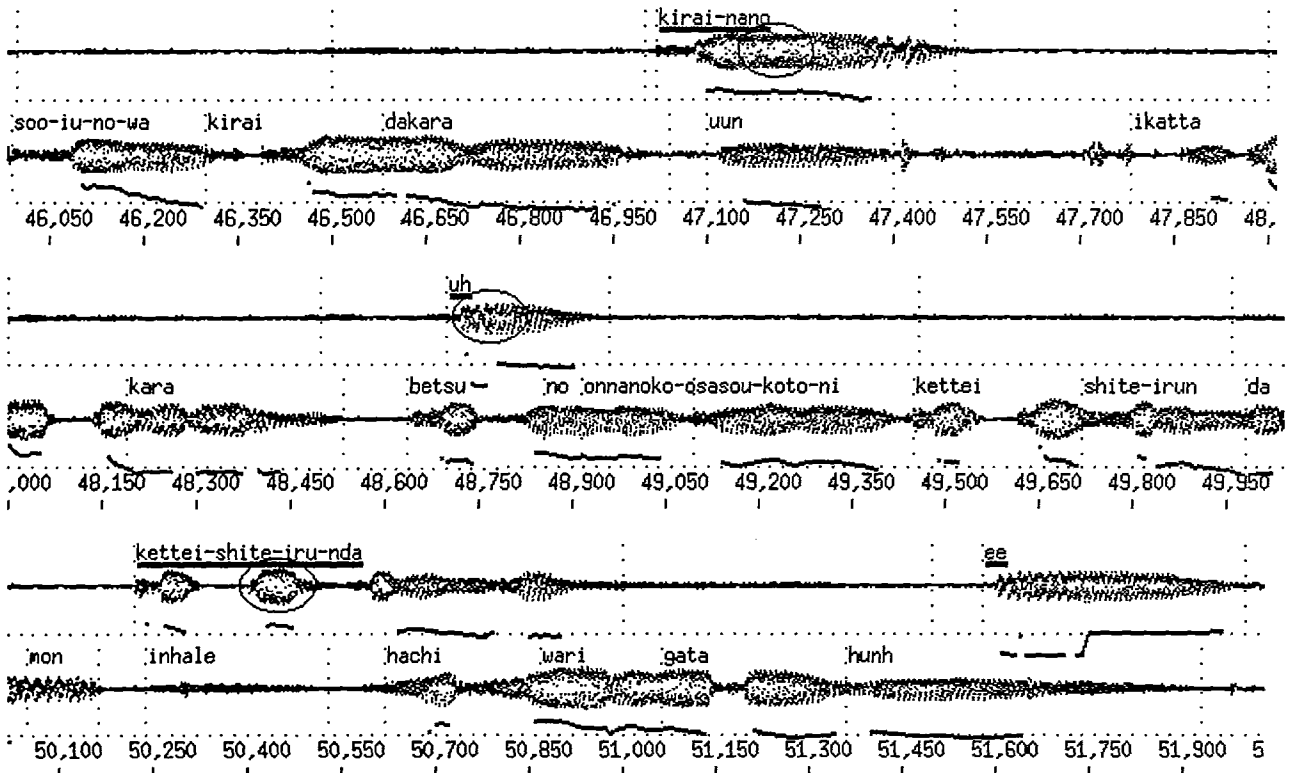


Figure 2: Conversation Fragment. Each of the three strips includes two rows, the top being conversant JH and the bottom being conversant KI. Each row includes: transcription, signal, pitch, and the 30th percentile pitch level (horizontal dotted line). Underlining indicates aizuchis, and the ovals indicate predicted aizuchi points; note that 3 out of 4 aizuchis were correctly predicted. The context is that tomorrow is KI's birthday, his girlfriend doesn't seem to be planning anything, and he has just rejected the suggestion that he forget his pride and just ask her to arrange something. A translation appears as Figure 3.

	<u>hate?</u>	
because I hate to do that.	really.	I'm pissed
	<u>mm</u>	
that's why.	I'll ask out some other girl, I've decided.	
	<u>you've decided!</u>	<u>well.</u>
inhale, I'm 80 percent sure.	you know.	

Figure 3: Free Translation of the Conversation in Figure 2

To better 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 aizuchis and rule-predicted aizuchis. Over the whole of this 5 minute conversation, compared to those produced by JH in response to KI, the rule correctly predicted 68% (54/79) of the aizuchis, 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 aizuchi.

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 stupid algorithm, 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 33% accuracy seems pretty poor, so we tried to characterize the error types. It turns out that over a third (roughly 343 of 933) of the incorrect predictions were at places where it seemed that an aizuchi would in fact have been appropriate. We believe that this reflects differences in aizuchi production style among speakers. Another 107 of the incorrect predictions were cases where the rule predicted an aizuchi 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, nor adaptive to speech rate etc.

3. Conversations with Humans

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

The key problem was how to get people to interact naturally with the system. The solution was to fool them into thinking they were interacting with a person. Hence I 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

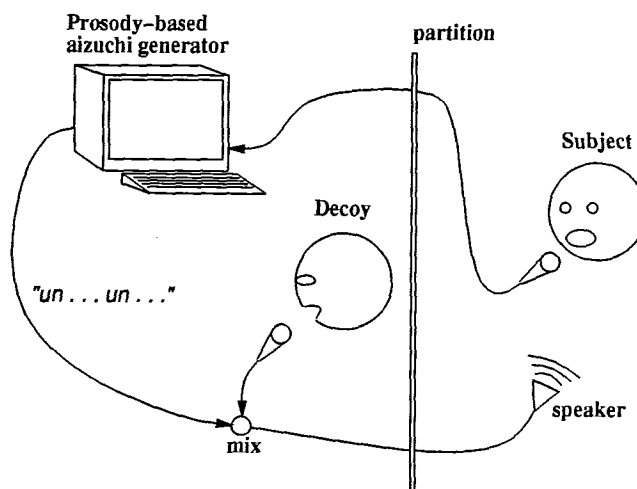


Figure 4: Experiment Set-up

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.
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 aizuchis (specifically the most neutral one, *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. 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. I am now trying to find a more sensitive metric of feedback quality; for example by doing longer runs and asking more penetrating questions, such as "in which case did you feel the other conversant was being more attentive/friendlier/more considerate/more understanding." Some ideas on how to achieve longer runs appear below.

4. Relevance of Embodied AI

To summarize how this system relates to the Embodied AI approach:

T2 By analogy to demonstrations of robots that interact with real unsimplified worlds, here we see that it is possible to build dialog systems that interact with real, unconstrained humans, not required to speak clearly, insert pauses, use only certain words, speak as if reading aloud, etc.

T3 By analogy to the idea that building a complete world model is impossible and unnecessary, we see that accurate recognition of all words and phonemes is not (always) necessary. More generally, the traditional concerns of the classic AI approach to dialog, cast as the problem of “speech understanding” (Erman *et al.* 1980) can be criticized in terms of methodology on many counts (Ward 1996a), many resembling points made in Brooks’ (1991) critique of the classic AI as a methodology for robotics.

T4 By analogy to behavior without use of a representation of the world, we see that direct linking of perception and action, without use of a representation of utterance meanings, is possible.

T5 By analogy to the appearance of intelligence emerging from interaction with the world, we see that 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 for dialog being responsive and keeping the interaction going may be more basic goals than being able to take dictation, answer questions about data in a table, etc.

T7 By analogy to the attention paid to ethology, we see the need for attention to the facts of human behavior. (Unfortunately, the work on language in social interaction, scattered among linguistics, psychology, sociology, and speech communication, is not quantitative, and psycholinguistics, which is quantitative, tends to focus on either comprehension in isolation or production in isolation. So it may be necessary to gather our own data and do our own analysis.)

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

Thus there seems to be a good match between the Embodied AI approach and a focus on the use of language in social interaction — in every respect except one: physical embodiment (T1) seems to be of secondary importance; not surprisingly, since it is of course common for people to interact verbally without seeing or touching each other. (Although of course gaze, nods, etc., are a necessary part of a full account of back-channel feedback (Duncan & Fiske 1985), let alone language in social interaction in general.)

5. Research Plans

So it seems that Embodied AI as a strategy and methodology provides a nice way to study language in social interaction, in at least one case, and probably in others. Indeed, related phenomena, including turn-taking phenomena, prosodically cued phenomena, and grunt-type responses, have already been targeted (Brooks & Stein 1993). This section mentions a few things high on the agenda for my students and I.

First, we plan to extend our system, retaining the basic format of responding passively and accepting any topic. We plan to 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 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 frequency, choice of word, and type of pronunciation for back-channel feedback.

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

Second, 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 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 control pacing etc (as done by Grunt (Schmandt 1994))
- on the other hand, being an appreciative listener, that is, producing appropriate back-channel feedback, for any of the above.
- participating in ritual conversations openings (“Hi Joe, how you doing? ...”)
- participating in naming and memory games (I-Spy-With-My-Little-Eye, or “Can you name all 50 states? I’ll keep track and give you hints if you get stuck ...”)
- 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 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) word(s) from each input. How to interface simple behaviors with an HMM-based recognizer is an open question.

Third, 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 — with lots of representations, lots of symbol manipulation, and a goal of 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 be helping the user with the task of solving the puzzle, this will be just a pretext for interacting naturally, and hopefully enjoyably, with the user.

Fourth, 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 human take. For example, we may be 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. Another direction is to look for evidence on the question of whether human language processing should be modeled with a subsumption architecture. Here we plan to look at cases of apparent dissociation among processes, including stuttering, speech errors, false starts and pauses (all of which are typically filtered out and not perceived by listeners, but are relatively common).

Finally, we plan to do all these things in several languages.

6. Summary

What we'd really like to do is build systems that use language (and learn language) in contexts like Duck-Duck Goose, Tag, and Cowboys and Indians. This being impractical, we have to settle for building disembodied language users.

The work described in this paper shows that we can exploit most of the tenets of the Embodied AI approach anyway, if we focus on the use of language for social interaction.

References

Brooks, Rodney A. (1991). Intelligence Without Reason. In *Proceedings of the Twelfth International*

Joint Conference on Artificial Intelligence, pp. 569–595.

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.

Erman, Lee D., Frederick Hayes-Roth, Victor R. Lesser, & D. Raj Reddy (1980). The Hearsay-II Speech-Understanding System: Integrating Knowledge to Resolve Uncertainty. *Computing Surveys*, 12:213–253.

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.

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). Second Thoughts on an Artificial Intelligence Approach to Speech Understanding. In *14th Spoken Language and Discourse Workshop Notes (SIG-SLUD-14)*, pp. 16–23. Japan Society for Artificial Intelligence. ftp: ftp.sanpo.t.u-tokyo.ac.jp/pub/nigel/papers/second-thoughts.ps.Z.

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

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