Using Prosodic Clues to Decide When to Produce Back-channel Utterances

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

University of Tokyo

ABSTRACT

Back-channel feedback is required in order to build spoken dialog systems that are responsive. This paper reports a model of back-channel feedback in Japanese dialog. It turns out that a low pitch region is a good clue that the speaker is ready for back-channel feedback. A rule based on this fact matches corpus data on respondents' production of back-channel feedback. A system based on this rule meets the expectations of live speakers, sometimes well enough to fool them into thinking they are conversing with a human.

1. MOTIVATION

Today's typical spoken dialog system produces no response until after the speaker finishes an utterance. Humans, in contrast, are very responsive, reacting frequently while the speaker is talking. Giving speech systems this ability may make interaction more pleasant and efficient (Johnstone *et al.* 1995). One important component of responsiveness is back-channel feedback, and a key question is when this is appropriate.

Japanese is a particularly interesting language in this regard, in that back-channel feedback occurs approximately twice as frequently as in English (Maynard 1989). It is such an essential part of dialog that Japanese has a non-technical term for instances of back-channel feedback: "aizuchi". This paper reports a basic study of back-channel feedback in Japanese, providing a partial answer to the question of when to produce such feedback.

2. DEFINITION

A workable definition is that back-channel feedback:

1. responds directly to the content of an utterance of the speaker,
2. is optional, and
3. does not require acknowledgement by the speaker.

These three characteristics distinguish back-channel feedback from some closely related phenomena: A. Characteristic 1 rules out speaker-produced grunts, which often seem to serve to emphasize the speaker's previous utterance. B. Characteristic 1 also rules out feedback which occurs several seconds after the speaker's utterance, seemingly reflecting the result of some cogitation. C. Characteristic 2 rules out grunts in response to questions. D. Characteristic 3 rules out questions, even *huh?*. E. Characteristic 3 also rules out feedback grunts which segue into full-fledged utterances. Of course, there is no clear boundary between back-channel feedback and these phenomena, and in perhaps 2% or 3% of the cases deciding whether something is back-channel feedback or not still feels arbitrary.

Characteristic 3 says "require" not "receive" because, although the speaker generally continues speaking after receiving back-channel feedback, this is not always the case. He may stop, or even respond explicitly to the feedback.

The existence of speaker-produced grunts (phenomenon A) raises a problem. These are often timed such that, if the respondent produces feedback for the previous utterance, the speaker-produced grunt directly follows the respondent's feedback and appears to be a response to it. Such grunts are impossible to distinguish from grunts that actually do respond to feedback. Erring on the side of caution, it seems best to not consider any grunts which respond to feedback to be back-channel feedback. In other words back-channel feedback should not count as "utterances" when applying characteristic 1.

This definition circumscribes roughly the same set of phenomena as other definitions in the literature. (Good entry points to the literature on back-channel feedback are Maynard (1989) and Novick and Sutton (1994).) It has however two advantages:

First, this definition is relatively easy to use to decide whether a specific utterance is back-channel feedback or not. This simplicity of application is obtained in part because it does not refer to function. This is appropriate because back-channel feedback has no consistent immediate effect, and the effects it does have, namely, effects on the flow of dialog over longer time frames (perhaps 2 to 20 seconds) are not directly relatable to specific instances of feedback. Another reason for the simplicity of application is that the definition does not refer to exactly what back-channel feedback is in response to. This is because feedback can respond to many things (including the mere fact of the other person speaking or trying to get started, as with grunts that serve to yield the floor after inadvertent simultaneous speaking, and also specific aspects of what the speaker expresses, such as facts, reasons, feelings, and referents), but in specific cases it is generally not possible to identify exactly what it responds to.

Second, this definition refers to back-channel feedback as a discourse phenomenon, rather than referring to the form, meaning, or length of the feedback. While there are words (actually short grunts) typically used for back-channel feedback, it seems a mistake to take their characteristics as definitional. Rather, back-channel feedback seems to encompass a continuum. At one extreme there is very subtle feedback, including laughter, coughs, sniffs, and barely audible grunts. In the middle there are grunts which are more clearly signals to the speaker but which still convey no semantic information. At the other extreme there is feedback which expresses interest, surprise, sympathy, approval, etc., echos a key word, or completes or restates the speaker's unfinished utterance.

3. RELATED RESEARCH

Back-channel feedback is not produced at random. Many researchers have speculated about the factors that determine when it is appropriate.

One likely factor is the expression of some new information by the speaker. This factor is popular among those who study imaginary conversations represented as text. It is also a major factor in staged conversations, where the participants are required to perform specific tasks and the exchange of information is made artificially important. However, in natural dialog the importance of information and meaning in invoking back-channel feedback is probably overrated.

Another type of likely factor is syntactic, such as completion of a grammatical clause.

The other class of likely factors is prosodic. The idea here is that the speaker provides some clues which tell the respondent when back-channel feedback is appropriate. One possible prosodic cue is simply the onset of silence at the end of an utterance. For Japanese, other prosodic factors suggested include a low pitch point (Sugito 1994), a slowing, volume increase, and pitch increase (Koiso *et al.* 1995), and a specific pitch contour (Okato *et al.* 1996).

4. CORPUS

To look for prosodic cues to aizuchi my students and I recorded 17 short Japanese conversations between pairs of university students, totaling 80 minutes. The instructions were basically just "We're studying aizuchis. Please have a conversation." Thus the conversations were unconstrained and natural. In most of the conversations the participants were seated in such a way as to prevent eye contact. Recording was done using head-mounted microphones in stereo onto DAT tape and the conversations were uploaded to a computer for labeling and analysis. By the definition of §2 this corpus includes 789 aizuchis.

A sample of a conversation from the corpus appears on the CD-ROM proceedings as sound [A062S01.WAV] and graphically, with aizuchi underlined [A062S01.GIF]. This figure also appears in (Ward 1996a).

5. PRELIMINARY ANALYSIS

For this corpus, none of the prosodic features mentioned in §3 seem to have a strong correlation with the appearance of aizuchis. In particular, the onset of silence at the end of an utterance cannot be the major cue. This is because it obviously can play no role for aizuchis which overlap the speaker's utterance or for aizuchis which follow the utterance end with a delay less than human reaction time, which is over 200ms — and such cases account for about two thirds of the aizuchis. By the same reasoning the length or volume of the last syllable or word of the utterance or phrase cannot be major factors.

6. PREDICTION RULE

In Japanese a region of low pitch means that back-channel feedback is appropriate.

More specifically, upon detection of the end of a region of pitch less than the 30th-percentile pitch level and continuing for 150ms, coming after at least 700ms of speech, you should produce an aizuchi 200ms later, providing you have not done so within the preceding 1 second. (The specific values here were obtained by tuning the parameters to get good agreement with the corpus.)

This rule is currently implemented as follows: First, energy is computed for each 10ms frame and a histogram of energy values is made. The lower peak in this histogram is considered the background energy level and the higher peak is considered the typical vowel energy level. Frames whose energy level is greater than (.8 × typical-vowel + .2 × background) are considered to be speech. For grouping speech frames into speech regions, gaps of up to 250ms of non-speech are allowed.

Second, the pitch is computed every 10ms, improbable values are discarded, and the distribution is computed. Frames with a pitch less the 30th percentile pitch level are considered to be low pitch frames. Frames at which no pitch was detected inherit the pitch of the most recent frame with a pitch, provided that frame was no more than 80ms away. This implies that gaps of less than 80ms are filled in. It also implies that a 70ms low pitch region at the end of an utterance counts as a 150ms low pitch region.

Conversations are handled as independent files of 1 minute each. This implies that the value of the 30th-percentile pitch is somewhat sensitive to pitch range variation, which is useful, for example, for handling increases in baseline pitch during interesting minutes of the conversation.

Clearly the details of this computation are ad hoc and could be improved in many ways.

7. CORRESPONDENCE WITH RESPONDENTS' PERFORMANCE

To evaluate the performance of the above rule, its predictions were scored as correct if the predicted aizuchi initiation point was within 500ms of that of an aizuchi produced

by the original human respondent. For some situations performance was very good. In particular, compared to the occurrences of aizuchis produced by JH in response to KI in their 5 minute conversation, the rule correctly predicted 69% (54/78), with an accuracy of 68% (54 correct predictions / 81 total predictions).

It is noteworthy that the rule handles both aizuchis which were produced after the speaker paused or stopped, and those which overlapped with his continued utterance.

It is also noteworthy that the rule handles both male and female speakers and respondents. (The only obvious difference between male and female aizuchi patterns is that with female-female pairs significantly longer aizuchis sometimes appear, for example *a-honto-ni–hee (oh, really, hmm)* lasting 1.3 seconds and *un-un-ee-ikitai (mm, mm, hmmm, I want to go)* lasting 1.5 seconds, neither of which caused the speaker to even pause. Such long aizuchis probably account for some of the "they're both talking at once and neither is listening" impression sometimes given by conversations among female friends.)

Running the rule on the entire corpus gave a coverage of 42% (333/789) and an accuracy of 25% (333/1342). For comparison, a random predictor's coverage was 18% (140/789) at an accuracy of 8% (140/1843).

Some ways in which the rule often fails are: 1. predicting an aizuchi where in fact the human respondent produced a near-aizuchi (mostly of types E, A, and D, as defined in §2), 2. predicting an aizuchi at every opportunity, whereas human respondents pass up about a third of the opportunities, 3. not predicting aizuchis which serve to mark yields. Most of the failures are more difficult to characterize.

The causes of the failures are diverse. Some of the failures are probably attributable to poor implementation and tuning of the rule – most obviously the lack of compensation for speaking rate. Most of the failures are probably due to factors not included in the rule. In particular, there is a clear need for: 1. dialog type factors (the rule does well for narrative and explanation, but not so well for banter, question and answer, instruction, teaching, ritual greetings, cooperative problem solving, and microphone tests), 2. prosodic factors other than low pitch, 3. semantic factors, and 4. factors involving dialect and personality of the speaker and respondent.

8. CORRESPONDENCE WITH SPEAKERS' EXPECTATIONS

I built a system to find out how well the above rule would perform in live conversation.

There were three critical issues. The first was how to compute pitch in real time. For this I used a a low sampling rate (8000 samples per second), and ran the pitch tracker on a fast machine (a Sun SparcStation 20). The second issue was how to produce appropriate aizuchis. It turned out to be acceptable to simply always produce *un*, the most neutral aizuchi. (In the corpus *un* was the most common aizuchi, accounting for 11% of the occurrences, and for 19% if variants like *uh*, *unn*, *hunn*, *hmm*, and *mm* are included.) Since always producing the same aizuchi sounded mechanical, I used two in alternation, or three with random selection. The third issue was how to get people to try to interact naturally with the system. The only 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 (see Figure 1). The aizuchis output by the system were recordings of decoy-produced samples, not synthesized. To make it impossible for subjects to distinguish between the decoy's live voice and the system's aizuchis, I introduced noise by over-amplifying both.

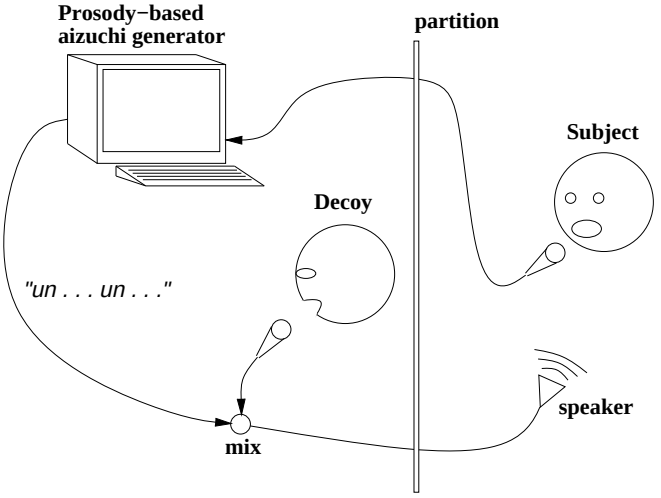

Figure 1: Experiment Set-up

The experimental 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 (eg, with "what project are you building in Mechatronics Lab this year?").
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.

I've done the experiment a couple of dozen times informally, as an exhibition at a symposium and also with whoever happens to visit the lab. In every case the system gives a strong impression of responding like a human. Many people don't notice anything unusual about the interaction.

I also did a more formal experiment, setting up things carefully to make it easier for the system. I used as decoy JH, the person whose conversational style the rule matched best. Also, to reduce the risk of subjects guessing the real purpose of the experiment, I used subjects who had previous

experience conversing with an unseen partner (specifically, in having contributed conversations to the corpus).

I did 4 runs, with different subjects. I used a slightly less accurate rule than that of §6. After switch-on the system contributed an average of 5.2 aizuchis and the decoy contributed an average of 5 utterances (including questions, answers, and aizuchis) over the course of a minute.

Afterwards I asked "was there anything strange about the conversation or about this person's (the decoy's) way of talking?". None of the subjects said yes, and all were surprised when told that their conversation partner had been partially automated. (This was ironic in that all the subjects were aware that I was trying to build system to fool people with aizuchis.) Thus it seems that the prediction rule produces aizuchis as speakers expect.

Of course, this result is probably due in part to a human tendency to be generous in interpreting a dialog partner's responses and response patterns, especially in real-time conversations.

9. SUMMARY

A low pitch region is an important cue for back-channel feedback production in Japanese. A rule based on this fact has been verified as matching respondents' feedback data and as meeting the expectations of live speakers.

10. SPECULATIONS

It is well known that prosody can express meaning or pragmatic force. What is new here is the evidence that prosody alone is sometimes enough to tell you what to say and when to say it. This confirms the intuition that you can often be responsive without paying attention to, let alone understanding, what is said to you. I imagine this is true not just for Japanese.

Thus the aizuchi-predicting rule discovered here is a "low-level behavior" in the sense that it involves a fairly direct link between perception and action. This suggests an analogy between the system of §8 and subsumption-based robots. This system interacts with a real human, doesn't think at all, and relies on a low-level behavior. Subsumption-based robots act in the real world, don't think too much, and rely on low-level behaviors (Brooks 1986). The analogy can be carried further. Since there seem to be other low-level behaviors in dialog, involving patterns of eye contact and patterns of what to pay attention to and how to react to it (Nagao & Takeuchi 1994; Ward 1996b), an appropriate model for combining dialog behaviors may be a "subsumption architecture" (Brooks 1986), where the various behaviors operate semi-autonomously and without central control. Such an architecture may be a good way to build a foundation for responsive and robust spoken dialog systems.

Acknowledgements

I thank Keikichi Hirose for the pitch tracker, Joji Habu for helping figure out how to fool subjects, Wataru Tsukahara for comments, and a couple dozen students for conversations and labeling.

References

Brooks, Rodney A. (1986). A Robust Layered Control System for a Mobile Robot. *IEEE Journal of Robotics and Automation*, 2:14–23.

Johnstone, Anne, Umesh Berry, Tina Nguyen, & Alan Asper (1995). There was a Long Pause: influencing turn-taking behaviour in human-human and human-computer dialogs. *Int. J. Human-Computer Studies*, 42:383–411.

Koiso, Hanae, Yasuo Horiuchi, Syun Tutiya, & Akira Ichikawa (1995). The acoustic properties of "subutterance units" and their relevance to the corresponding follow-up interjections in Japanese. (in Japanese). In *AI Symposium '95 (SIG-J-9501-2)*, pp. 9–16. Japan Society for Artificial Intelligence.

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.

Novick, David G. & Stephen Sutton (1994). An Empirical Model of Acknowledgement for Spoken-Language Systems. In *Proceedings 32nd Association for Computational Linguistics*, pp. 96–101.

Okato, Yohei, Keiji Kato, Mikio Yamamoto, & Shuichi Itahashi (1996). Prosodic pattern recognition of insertion of interjectory responses and its evaluation. (in Japanese). In *10th Spoken Language Information Processing Workshop Notes (SIG-SLP-10)*, pp. 33–38. Information Processing Society of Japan.

Sugito, Miyoko (1994). *Nihonjin no Koe.* Izumi Shoin.

Ward, Nigel (1996a). In Japanese a Low Pitch Region means "Backchannel Feedback Please". In *11th Spoken Language Information Processing Group Workshop Notes (SIG-SLP-11)*, pp. 7–12. Information Processing Society of Japan. ftp: ftp.sanpo.t.u-tokyo.ac.jp/pub/nigel/papers/low-pitch.ps.Z.

Ward, Nigel (1996b). Reactive Responsiveness in Dialog. In *AAAI Fall Symposium on Embodied Cognition and Action.* (submitted).