In Japanese a Low Pitch Means "Back-Channel Feedback Please"

Nigel Ward
University of Tokyo

Back-channel feedback is required to make spoken dialog systems that are responsive. In Japanese, 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' aizuchi production. A system based on this rule meets the expections of live speakers, sometimes well enough to fool them into thinking they are conversing with a human.

発話の中にピッチが低い領域があったらあいづちを打つ

N. ワード
東京大学

反応性の良い音声対話システムを作るには、会話の途中であいづちを打つ機能が必要である。日本語においては、話者の低ピッチ領域の生成に対して聞き手があいづちを打っていることが分かった。この規則に基づいてあいづちの発生を予測すると、人間同士の自然な会話データにおけるあいづちと良く一致した。そこでこの規則に従ってあいづちを打つシステムを構成し、被験者に会話させたところ、相手を人間だと思わせることができた場合もあった。

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 speaking. Giving speech systems this ability may make interaction more pleasant and efficient. One important component of responsiveness is back-channel feedback. This paper reports a model of back-channel feedback in Japanese dialog. (Henceforth I will use the term "aizuchi" as shorthand for "back-channel feedback in Japanese.")

2 Corpus

To study aizuchi my students and I recorded 17 short Japanese conversations between pairs of university students, totaling 80 minutes. The instructions were basically "We're studying aizuchi. Please have a conversation for 5 minutes." 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. A sample appears in Figures 1 and 2.

Conversations were recorded with head-mounted microphones in stereo onto DAT tape and uploaded to a computer for analysis.

3 Definition of Aizuchi

After many hours spent on the corpus, I arrived at some guidelines for what to label as an aizuchi. These correspond in general to most native speaker's intuitions, and are relatively easy to apply. Reducing the guidelines further led to a working definition: An aizuchi:

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

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

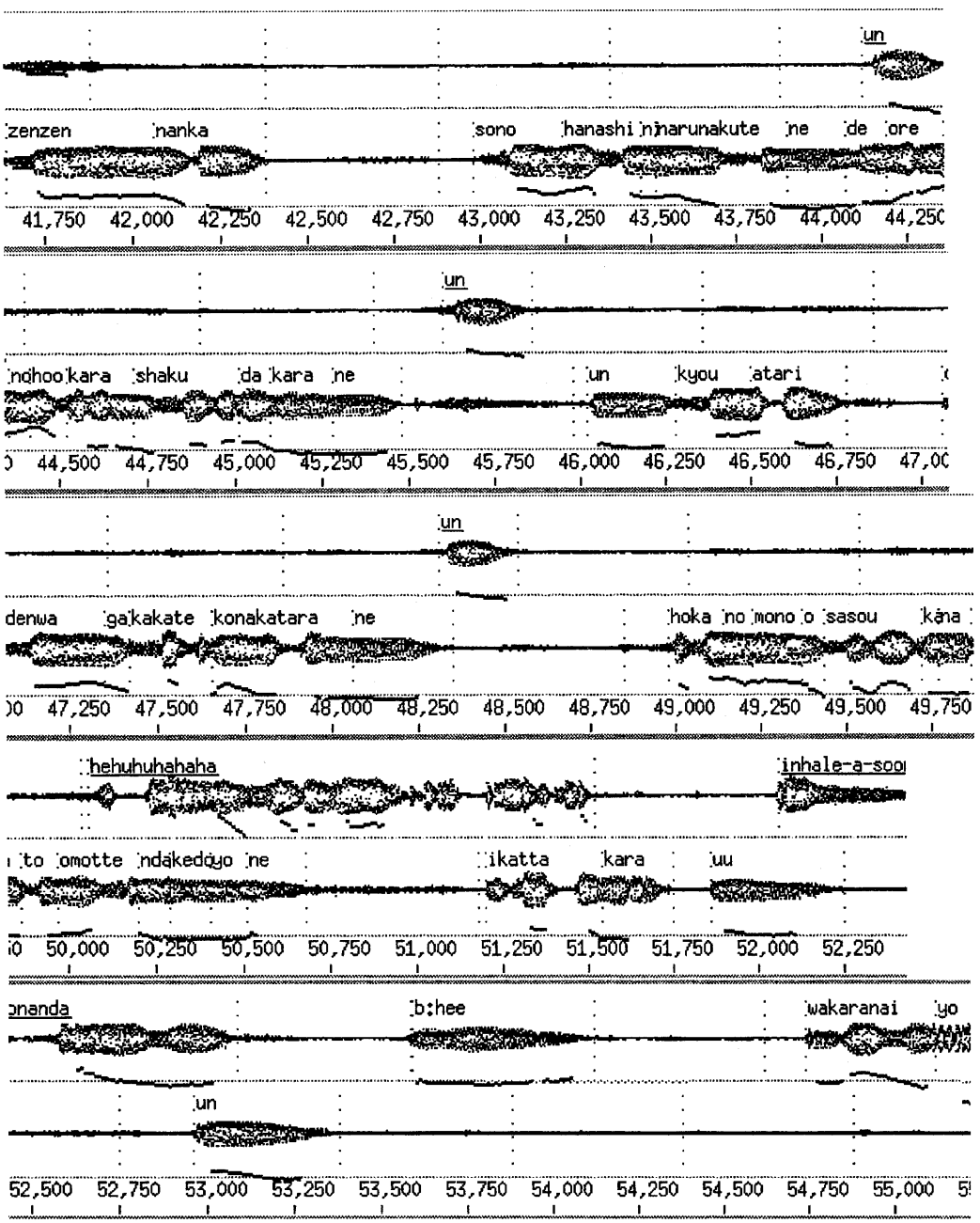

Figure 1: Conversation Fragment. The rows are: transcription with aizuchis underlined, signal, pitch, and low-pitch threshold (dotted); for conversants JH and KI. The context is that tomorrow is KI's birthday but his girlfriend doesn't seem to be planning anything. Noise from other conversations is evident around 41700ms and 45600ms.

These three characteristics distinguish aizuchis from some closely related phenomena: A. 1 rules out speaker-produced grunts, for example at 52000ms in Figure 1, which often seem to serve to emphasize the speaker's point. (These are often timed such that, if the respondent produces an aizuchi for the previous utterance, the speaker-produced grunt directly follows the aizuchi and appears to be a response to it. Such grunts are impossible to distinguish from grunts that do respond to an aizuchi, so, for consistency, I considered all aizuchi-acknowledging grunts (for example, at 53000) to not be aizuchis.) B. 1 also rules out aizuchi-like grunts which occur several seconds after the speaker's utterance, seemingly reflecting the result of some cogitation, for example at 53600ms. C. 2 rules out grunts in response to questions. D. 3 rules out questions, even *un?* (meaning *huh?).* E. 3 also rules out aizuchi-like grunts which segue into a full-fledged utterance, for example, at 46100ms (which is also an example of phenomenon A). Of course, there is no clear boundary between aizuchis and these phenomena, and in 2% or 3% of the cases, deciding whether something is an aizuchi or not still feels arbitrary.

Note that characteristic 3 says "require" not "receive". This because, although the speaker generally continues speaking after receiving an aizuchi, this is not always the case. An example is at 52100, which is counted as a aizuchi even though the speaker responds to it and stops.

> Somehow the subject never comes up, you know, *mm*, and I'm not going to beg, *mm*, and, by today, if she doesn't call me, you know, *mm*, I'll ask out some other girl, is what I'm thinking, *laughs*, you know, because it makes me mad, *(inhale) hmm*, um. *Hmm. I don't know ...*

Figure 2: Free Translation of the Conversation Fragment in Figure 1

This working definition does not refer to the function of aizuchis. This is unavoidable, since they have no consistent immediate effect; the effects they do have, namely, effects on the flow of dialog over longer time frames (perhaps 2 to 20 seconds), are unfortunately not usable as criteria for deciding how to label individual aizuchis.

A good entry point to the literature on issues in the definition of aizuchis is Maynard (1989).

Within the class defined by these three characteristics, there is great variety. One dimension of variety is the richness of the aizuchi. The prototypical aizuchi is a grunt, conveying little semantic information but a clear signal to the speaker. There are also more subtle aizuchis, including laughter, coughs, sniffs, and barely audible *uns*. On the other hand, there are also more expressive aizuchi, where the respondent expresses interest, surprise, sympathy, or approval, echos a key word, or completes or restates the speaker's unfinished utterance. Such aizuchi can be fairly long and complex; 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 pause. (Such long aizuchis seem to occur only in conversations between women; they may account for some of the "they're both talking at once and neither is listening" impression sometimes given by conversations among female friends.) Another dimension of variation is what exactly the aizuchi responds to. In the corpus there are aizuchi that refer to the fact of the other person speaking or trying to get started; in particular, those that serve to yield the floor after inadvertent simultaneous speaking. There are also aizuchi which refer to the specific aspects of what the speaker expresses, including facts, reasons, feelings, and referents.

In the corpus there were 789 aizuchis.

4 Related Research

Aizuchi are not produced at random. Many have speculated about the factors that determine when an aizuchi 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 as an aizuchi-affecting factor is probably overrated.

The other class of likely factors is prosodic. The idea here is that the speaker provides some clues which tell the respondent when an aizuchi is permissible.

One possible prosodic clue is simply the onset of silence at the end of an utterance. However, this cannot be a factor for aizuchis which overlap the speaker's utterance, or for aizuchis which follow the phrase end with a delay less than human reaction time, which is over 200ms — and such cases account for most of the aizuchis. For the same reason, the length or volume of the last syllable or word of the utterance or phrase cannot be major factors.

For Japanese in particular, 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). In my data, none of the above seemed to have a strong correlation with the appearance of aizuchis.

5 An Aizuchi Prediction Rule

A region of low pitch means that an aizuchi 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, providing 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 effectively 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, for example, when baseline pitch increases during interesting minutes of the conversation.

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

6 Correspondence with Respondents' Performance

To evaluate the performance of the above rule, 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 aizuchis produced by JH in response to KH in the 5 minute conversation from which Figure 1 was taken, the rule correctly predicted 69% (54/78) of the aizuchis, with an accuracy of 68% (54 correct predictions / 81 total predictions). In particular, the aizuchis at 44200, 45600, and 48300 were predicted, the aizuchis at 50100 and 52100 were missed, and there was an incorrect prediction at 50800.

It turns out that the rule handles both

aizuchis which were produced after the speaker paused or stopped, and those which overlapped with his continued utterance.

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 §3), 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. However 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, instructions, 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.

7 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 (8K samples per second), and ran the pitch tracker on a fast machine (a Sun SS20). 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 3). 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.

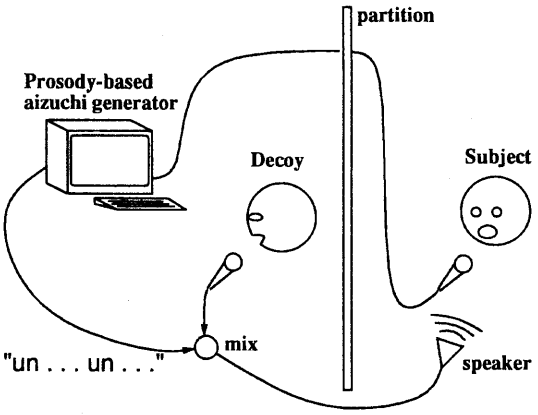

Figure 3: Experiment Set-up

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). (Although all the subjects were aware that I was planning to build a system to fool people with aizuchis, none were suspicious about the set-up.)

I did 4 runs, with different subjects. I used a slightly less accurate rule than that of §5. 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. 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.

8 Summary

A low pitch region is an important cue for aizuchi production. A rule based on this correlation has been verified as matching respondents' aizuchi data and as meeting the expections of live speakers.

9 Larger Significance

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 is arguably the first demonstration of a direct link between perception and action in language use. This suggests a subsumption approach to the construction of speech dialog systems (Ward 1996).

Acknowledgements: Keikichi Hirose gave me the pitch tracker. Joji Habu helped me figure out how to fool subjects. Wataru Tsukahara commented on the paper. A couple dozen students contributed as data sources, as labelers, and as subjects.

References

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*, pp. 9–16. Japan Society for Artificial Intelligence. SIG-J-9501-2.

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

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

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

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