# Some Exotic Discourse Markers of Spoken Dialog

Nigel Ward[1]
University of Tokyo

**Abstract:** In Japanese and English conversation corpora we have found an interesting discourse marker: a 110 millisecond region of low pitch region, which can cue back-channel feedback from the hearer. We have also found that many back-channel responses appear to be more explainable if considered not as words, but as complex sounds which directly convey discourse functions with sound symbolism.

## 1 Introduction

Back-channel feedback, also called "listener responses", is, to a first approximation, those responses produced by one participant which do not interfere with utterances by the other participant. In American English *yeah*, *mm* and *uh huh* are typical back-channel feedback. In Japanese *un* is most typical.

This paper summarizes some findings on back-channels and their cues, as primitive examples of discourse markers in spoken dialog. Definitions, details, discussion, and references appear elsewhere (Ward 1996; Ward 1998; Ward & Tsukahara to appear: Ward submitted; Ward & Tsukahara 1998). The findings arose from our corpora of unrestricted two-person conversations in Japanese and in American English, and might not apply to task-directed discourses, small talk, multi-party conversations, and conversations with audiences.

## 2 A Cue for Back-channel Feedback

In Japanese and in English, the prosody of the speaker's utterances can mark times when the listener is welcome to produce back-channel feedback. One specific cue is a region of low pitch. Behavior of Japanese listeners can be modeled with the following rule:

Upon detection of
a) a region of pitch less than the 28th-percentile pitch level and
b) continuing for at least 110ms,
c) coming after at least 700ms of speech,
d) providing you have not output back-channel feedback within the preceding 1.0 seconds,
e) after 350ms wait
you should produce back-channel feedback.

Our model for English is the same except for some parameters: pitch level (clause a) 26th-percentile, recovery time (clause d) 800ms, and delay (clause e) 700ms.

## 3 Corpus-based Evaluation

We have tested the predictions of the above rule against a corpus of casual, mostly friendly Japanese conversations between mostly students (80 minutes total).

The rule gave a coverage of 56% (of 873 back-channels in the corpus, the rule predicted 496, using a 500 ms tolerance) and an accuracy of 34% (of 1447 predictions, 496 were correct). This did better than a rule which makes predictions at random (while obeying clauses c, d, and e), both on average (Table 1), and for most speakers in most conversations: in 34 cases out of 36 (2 sides times 18 conversations) the figure of merit (namely the product of coverage and accuracy) was higher for the low pitch rule.

Analysis of the false predictions for Japanese suggests that roughly half are due to inter-speaker differences in back-channel behavior; this is the source of the "eavesdropping" estimate.

Similar results were obtained for a corpus of 68 minutes of English conversation data for the English rule (Table 2), although the accuracy was much lower, suggesting that in English factors other than low pitch are relatively more important than in Japanese.

## 4 Informal Experiments

We have also tested the performance of the rules in live conversation, mostly over the telephone. We used a human decoy to start the conversation. After the subject started talking, the decoy shut up and the system took over, producing back-channel feedback in response to regions of low pitch. This feedback was *un* for Japanese and a random selection between *uh-huh* and *mm* for English. We used pre-recorded samples of the decoy's voice to make it impossible for

[1] nigel@sanpo.t.u-tokyo.ac.jp, http://www.sanpo.t.u-tokyo.ac.jp/~nigel/, Mech-Info Engineering, University of Tokyo, Bunkyo-ku Tokyo 113-8656 Japan.
Thanks to Wataru Tsukahara, Dan Jurafsky, and anonymous reviewers for suggestions, and to the Hayao Nakayama Foundation and the Inamori Foundation for support.

| Predictions from | Coverage | Accuracy | Figure of Merit |
|---|---|---|---|
| low pitch regions | 56% (496/873) | 34% (496/1447) | .195 |
| random | 25% (222/873) | 24% (222/915) | .062 |
| utterance end | 68% (593/873) | 22% (593/2751) | .146 |
| utterance end and low pitch region | 36% (314/873) | 32% (277/978) | .115 |
| utterance end and no low pitch region | 32% (279/873) | 16% (279/1773) | .050 |
| eavesdropping human judge (estimate) | 95% | 67% | .64 |

Table 1: Performance of Various Rules for Predicting Back-channel Feedback (Japanese)

| Predictions from | Coverage | Accuracy | Figure of Merit |
|---|---|---|---|
| low pitch regions | 48% (172/359) | 18% (172/936) | .088 |
| random | 22% (80/359) | 13% (80/618) | .029 |
| utterance end | 46% (164/359) | 10% (164/1698) | .044 |
| utterance end and low pitch region | 30% (109/359) | 19% (109/578) | .057 |
| utterance end and no low pitch region | 15% (55/359) | 5% (55/1120) | .008 |

Table 2: Performance of Various Rules for Predicting Back-channel Feedback (English)

subjects to distinguish between the decoy's live voice and the system's output. If the conversation flagged, the decoy would speak up with another question or comment to get the subject talking again.

We have done a few dozen runs in both Japanese and English. In general third party judges listening to the conversations could distinguish the low pitch based aizuchis from randomly produced ones; the former sounded natural and the latter sounded odd, with clear cases of inappropriate aizuchis and of inappropriate silences when an aizuchi was called for. However, those who were actually talking to the system were apparently seldom aware of nor affected much by when or whether back-channels were produced.

## 5 Communicative Functions

The 110ms low pitch region has no single fixed meaning or function, at least in our data.

One thing it often co-occurs with is completion of a grammatical clause. Here it often seems to serve as an indication that the speaker considers that he has transmitted some new information, and so the hearer is welcome to confirm receipt or understanding or interest, with a back-channel. (We can think of it as conveying "this completes that thought, did you follow?") Sometimes what has been transmitted is a complete new fact or proposition, but often it is the introduction of just enough information for the listener to infer the speaker's point, especially in Japanese. In such cases back-channel feedback can appear before the speaker has completed a grammatical phrase or logical proposition, and sometimes back-channel feedback in such cases takes the form of completing the speaker's thought or sentence.

The low pitch region also often co-occurs with repetitions of a word previously spoken, produced for emphasis or clarity and/or when recovering from a false start, especially in English. In such cases it often welcomes back-channel feedback, perhaps conveying "I said it again, did you get it that time?"

The low pitch region also occurs frequently with disfluencies and markers of formulation difficulties, especially in English. In these cases we can imagine that the low pitch region conveys, "I'm stuck, but keep listening, something meaningful will come out soon". It also occasionally occurs as a speaker takes the floor. In some of these cases, especially for Japanese, it elicits back-channel feedback, presumably as encouragement to continue.

Another place where the low pitch region often occurs is together with back-channel feedback itself. Such cases of back-channel feedback themselves occasionally elicit a confirmatory word or sigh.

## 6 Co-occurring Markers

The low pitch cue tends to occur together with other discourse markers.

It co-occurs frequently with specific lexical items, as one would expect from the communicative functions identified in the previous section. In Japanese it often occurs with clause connectives (most commonly *kara*, *-te* and *kedo*), with 'agreement seeking sentence-final particles', especially *ne*, and with the back-channel *un*. In English the association with specific lexical items is less strong, but the low pitch region falls most frequently on *the* (almost always in the lengthened, unreduced pronunciation indicating difficulty finding the next word to say), *and*, and *um*.

The low pitch region is often followed by silence at the end of a speaker's utterance. The energy drop that marks the start of this silence is, counter-intuitively, not much of a cue for back-channel feedback, providing little or no information beyond that provided by the low pitch cue, as seen in the tables (results are for a rule which predicts a back-channel in response to 150ms of silence, subject to clauses c, d, and e of the corresponding low pitch rule). This also implies that the low pitch region is often a valid cue even when it appears in the middle of an utterance.

The low pitch region occasionally segues into a rise in intonation (uptalk). This seems to turn an invitation for feedback into a demand for it.

The low pitch cue sometimes co-occurs with vowel lengthening. This may be a consequence of the need to produce a low pitch region of sufficient length, in those cases where there is only a single syllable of lexical content to work with, for example with *ne* ('you know').

Gaze, posture, and facial and hand gestures also may correlate with low pitch regions.

Given all these correlations, it is natural to wonder whether it is necessary to invoke a notion of low pitch cue to explain the data. We have found that no other single factor can for all the occurrences of back-channels that low pitch regions can; to say more will require further analysis.

## 7 Responses Evoked

There have been many studies of the lexical items used in back-channels and the types of semantic functions served thereby.

This section discusses one problematic subset of back-channels, those sounds which do not seem to be words. For example, in the Japanese corpus, in addition to the ubiquitous *un*, there is also *uu*, *uh*, *uun*, *ununun*, *huun*, *huh*, *hmmm*, *hm-um*, and over a hundred other items not found in dictionaries, with diverse prosody and voicing. In English, there is a family containing *uh-huh*, *um-hm*, *uh-hm*, *hm*, *hmm*, *mm*, *un*, and *ahhh*, and another containing *okay*, *kay*, and *n-kay*.

Rather than consider each of these a distinct lexical item, a more parsimonious account may be reached by means of the following hypotheses: A: these sounds are not fixed sequences of phonemes, but are formed for each occasion from basic acoustic components; B: these acoustic components individually bear meanings; C: the meaning of a combination of acoustic components is the combination of the meanings of each component.

Some specific hypothesized meanings, for Japanese and possibly English too, are agreement for nasalization, contemplation for *m*, deference for breathiness and *h*, willingness to listen for number of syllables, and coldness for sharpness of final energy drop. In addition energy and pitch height and slope appear to bear the usual meanings.

If these hypotheses are correct, then these conversational sounds are 'iconic', or, in other words, involve 'sound symbolism' or 'synaesthesia'. This discourse-functional system of sound symbolism appears to be distinct from the onomatopoeic and mimetic systems of sound symbolism.

## 8 Envoi

Examination of simple markers for some cross-speaker discourse relations in spoken dialog has shown that prosody plays an important role. An important next topic is study of the interplay between prosodic discourse markers and lexical discourse markers.

# References

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

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 (submitted). On Back-Channel Feedback and Cooperation in Spoken Dialog. *International Journal of Human-Computer Studies*.

Ward, Nigel & Wataru Tsukahara (1998). Prosodic Features which Cue Back-Channel Feedback in English and Japanese. manuscript.

Ward, Nigel & Wataru Tsukahara (to appear). A Responsive Dialog System. In Yorick Wilks, editor, *Machine Conversations*. Springer Verlag.