ELSEVIER

Journal of Pragmatics 32 (2000) 1177–1207

journal of PRAGMATICS

www.elsevier.nl/locate/pragma

Prosodic features which cue back-channel responses in English and Japanese☆

Nigel Ward*, Wataru Tsukahara

Mechano-Informatics, Engineering, University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo 113-8656, Japan

Received 11 August 1998; revised version 23 August 1999

Abstract

Back-channel feedback, responses such as *uh-uh* from a listener, is a pervasive feature of conversation. It has long been thought that the production of back-channel feedback depends to a large extent on the actions of the other conversation partner, not just on the volition of the one who produces them. In particular, prosodic cues from the speaker have long been thought to play a role, but have so far eluded identification. We have earlier suggested that an important prosodic cue involved, in both English and Japanese, is a region of low pitch late in an utterance (Ward, 1996). This paper discusses issues in the definition of back-channel feedback, presents evidence for our claim, surveys other factors which elicit or inhibit back-channel responses, and mentions a few related phenomena and theoretical issues. © 2000 Elsevier Science B.V. All rights reserved.

Keywords: Back channel; Feedback; Cue; Prosody; Low pitch; Continuer

1. Introduction

By 'back-channel feedback' we mean, approximately, the short utterances produced by one participant in a conversation while the other is talking. In English, typical expressions used in back-channel feedback include *yeah*, *uh-huh*, *hm*, *right* and *okay*. Back-channel feedback is pervasive in conversations in English and in Japan-

☆ We thank Keikichi Hirose for the pitch tracker, Yuichiro Fukuchi for efforts to disprove our hypothesis, Daniel Jurafsky, Kikuo Maekawa, Elizabeth Shriberg, Maki Sugimoto, Minoru Terada and anonymous referees for discussion, and our conversants for the data. We also thank the Japanese Ministry of Education, the Sound Technology Promotion Foundation, the Nakayama Foundation, and the Inamori Foundation for support.

* Corresponding author. Phone: +81 3 5841-6346; Fax: +81 3 3 818 0835 E-mail: nigel@sanpo.t.u-tokyo.ac.jp

0378-2166/00/$ – see front matter © 2000 Elsevier Science B.V. All rights reserved.
PII: S0378-2166(99)00109-5

ese, the two languages addressed by this paper. One analysis of 1155 American English conversations found 19% of the utterances to be back-channels (37,096 out of 205,000 total utterances) (Jurafsky et al., 1997). Back-channels seem to be even more frequent in Japanese (Maynard, 1989).

Many researchers, working from diverse perspectives, have considered back-channel feedback, along with other turn-taking phenomena, to be a phenomenon of special interest, as being prototypical of social interaction in general (Yngve, 1970; Sacks et al., 1974; Duncan and Fiske, 1985; Ward, 1997b). In particular, there is the mystery of how 'coordination' is achieved – when two people are talking together, their utterances seldom interfere with each other, despite the lack of any fixed protocol for who may speak when.

Many researchers have sought for an answer in terms of 'signals'. For back-channel feedback in particular, it has long been thought that there may be a signal that tells a listener that 'it's now appropriate to respond with back-channel feedback', and that this signal would be prosodic, rather than involving meaning (Yngve, 1970). Previous research has discussed such possible signals as low pitch points, lengthening, and 'non-final intonation contours', as detailed in Sections 3.1, 5.1 and 5.2.

This paper reports the tentative identification of low pitch regions as signals in English and in Japanese which 'cue' back-channel feedback. Section 2 proposes a definition of back-channel feedback. Section 3 presents a prosodic feature, a region of low-pitch, that predicts the subsequent occurrence of back-channel feedback. Section 4 provides corpus-based evidence for the relevance of this feature. Section 5 discusses other factors relating to these prosodic features and to back-channel feedback. Section 6 discusses some open issues.

2. (Issues in the) Definition of back-channel feedback

This section delimits the phenomenon of back-channel feedback as studied in this paper, and discusses its correlates, its sub-types, and related phenomena, in English and in Japanese. The definition developed draws on previous surveys and discussions of various proposed definitions (Rosenfeld, 1978; Schegloff, 1982; Oreström, 1983; Mizuno, 1988; Maynard, 1989; Drummond and Hopper, 1993; Clancy et al., 1996; Horiguchi, 1997). Most of the points we make reflect a majority of opinion among researchers, although there are divergent opinions on each issue.

2.1. Clear cases

Clear cases of back-channel feedback happen somewhat like this: One person is explaining something or telling a story, the other person is paying attention and understanding, and produces a typical word or sound to indicate this, and also to indicate that he wishes the story-teller to continue. The story teller, without showing any awareness of this response, continues with his story, perhaps slightly encouraged to know that his listener is still interested.

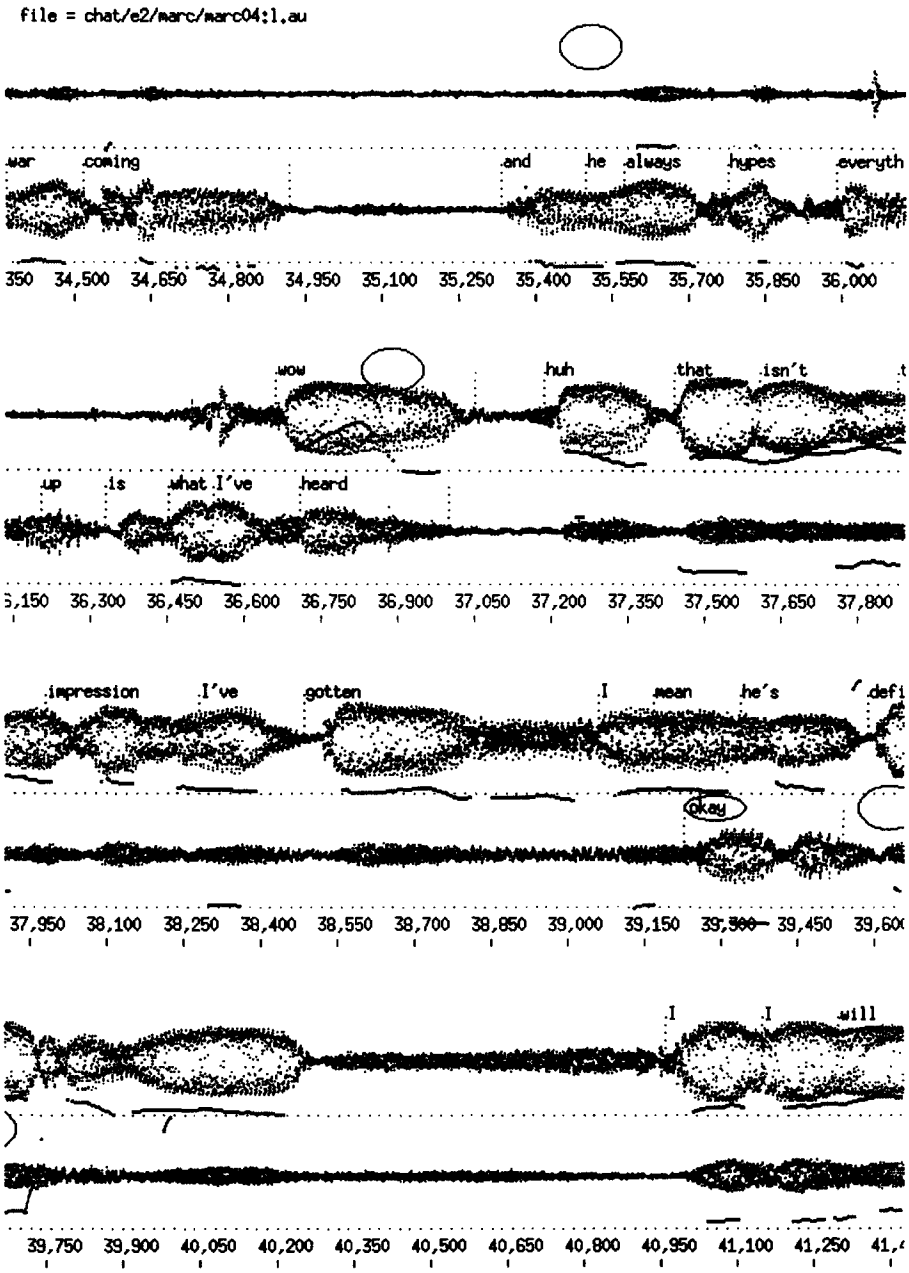

Fig. 1. English conversation fragment. Each of the four strips includes two tracks and a timeline. In each strip the top track is one speaker and the bottom track the other. Each track includes: a transcription, the signal, the pitch, and the 26th percentile pitch level (horizontal dotted line). Narrow ovals indicate actual back-channel feedback. Wide ovals indicate predicted back-channel feedback (section 3.2). The context is that the bottom speaker has complained that a certain newspaper reporter is biased; the top speaker has a different impression.

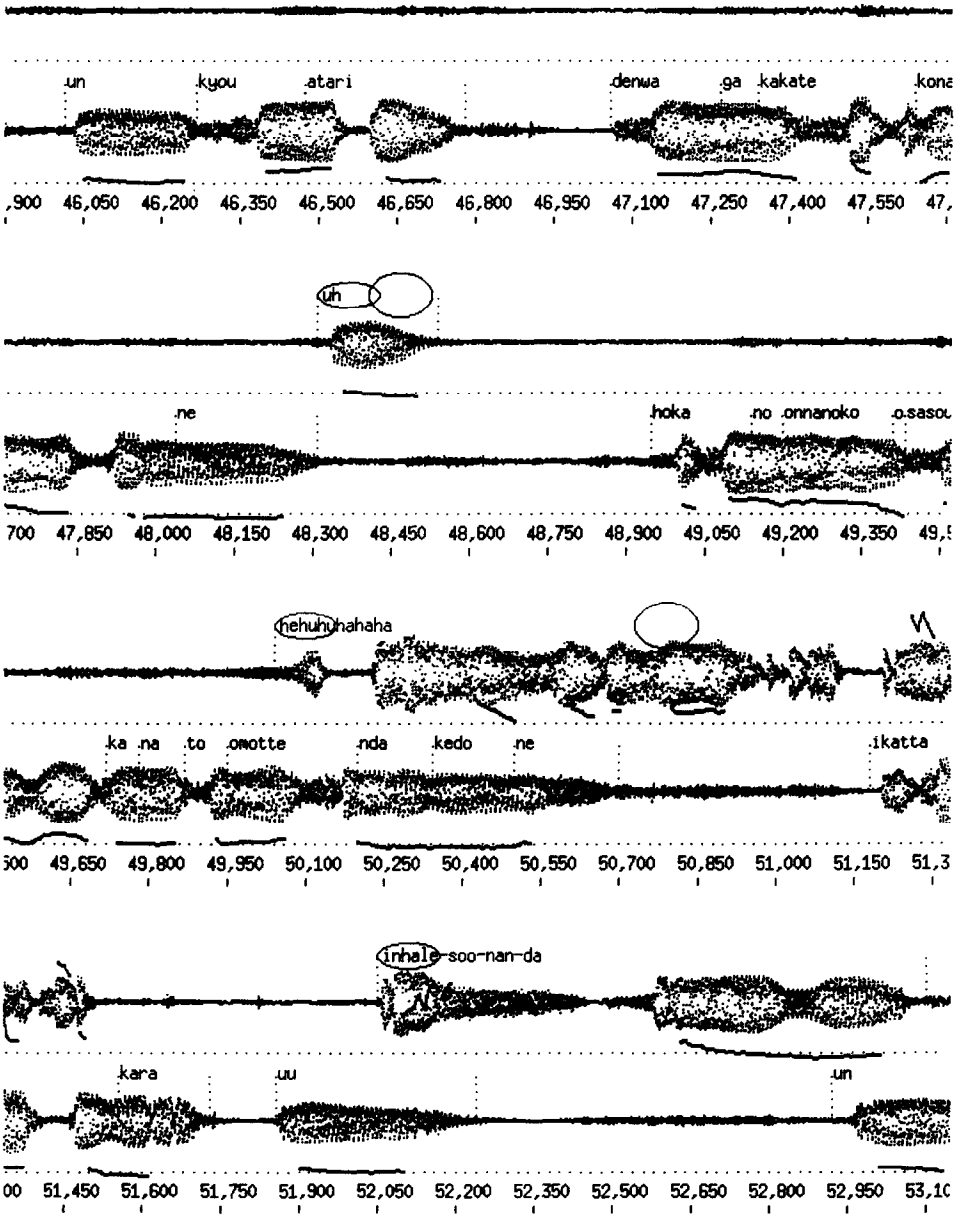

Fig. 2. Japanese conversation fragment in the same format as the previous figure, except that here the horizontal dotted line indicates the 28th percentile pitch level. The context is that the bottom speaker is worried that his girlfriend doesn't seem to be planning anything for his birthday.

Fig. 1 shows such an English conversation fragment containing back-channels and Fig. 2 shows a Japanese fragment, with Fig. 3 providing a gloss and Fig. 4 a rough translation.

hmm today around		telephone SUBJ call not-come		
you-know	mm	another girl OBJ ask-out		
QUEST hmm QUOT think			[laughter] PROG but you-know	angry
because um	really	um		

Fig. 3. Gloss of Figure 4

hmm today,		if there's no telephone call from her,		
you-know	mm	I'll ask out some other girl,		
is what I'm thinking.			[laughter] you-know	I'm mad
that's why. um		really	um	

Fig. 4. Translation of Fig. 4.

These clear cases are the typical examples that all researchers in this area use when evoking what sort of phenomena they are setting out to study (Yngve, 1970; Schegloff, 1982). They also seem to comprise much of the phenomena that have been studied as 'listener responses', 'accompaniment signals', 'continuers', 'assessments', 'acknowledgments', 'reactive tokens', 'interjectory utterances', and 'recipiency tokens'. These clear cases also correspond well to pre-theoretical notion of *aizuchi*, a common Japanese term. Since clear cases account for about 90% of what we eventually decided to give the label 'back-channel feedback', the exact definition

below is not crucial to the conclusions of the paper. For the sake of comparability with previous work, however, we do give details below.

2.2. Our definition

We arrived at our definition by starting from the clear cases, adding in cases similar to them, and trying to find a principled way to describe the expanded category. We refined our definition several times, striving for consistency and to produce simple and unambiguous guidelines usable by everyone involved with the corpus. Eventually we arrived at the following working definition:

Back-channel feedback:
(D1) responds directly to the content of an utterance of the other,
(D2) is optional, and
(D3) does not require acknowledgement by the other.

Note that this definition focuses, not on how these utterances fit into the structure of the discourse, nor on how they are evoked or perceived by the other, but instead on the perspective of the person producing them.

2.3. Related phenomena

The three clauses of our definition serve to distinguish back-channel feedback from some closely related phenomena: D1 rules out 'post-completion' vocalizations, produced by the speaker who has just produced an utterance, for example *uu* at 51,900 in Fig. 2. D1 also rules out feedback which occurs several seconds after the speaker's utterance, seemingly reflecting the result of some cogitation. D2 rules out responses to questions. D3 rules out most questions, including requests for clarification, such as *huh?*. (However it does not rule out questions which do not seek to interrupt or redirect the speaker, such as some occurrences of *sore de?* (and then?) in the Japanese corpus. These are rare.) D3 also rules out feedback sounds which segue without pause into full-fledged utterances.

Of course, there is no clear boundary between back-channel feedback and these phenomena. In perhaps one percent of the cases, deciding whether something in the corpus was back-channel feedback or not still felt arbitrary, even after both labelers listened together and discussed it.

The existence of post-completion vocalizations raises a minor problem. These are sometimes timed such that, if the respondent produces feedback for the previous utterance, the post-completion vocalization directly follows the respondent's feedback and appears to be a response to it, as at 52,950 in Fig. 2. Such vocalizations are impossible to distinguish from vocalizations that actually do respond to feedback. Erring on the side of caution, we do not consider any such vocalizations to be back-channel feedback. In other words back-channel feedback does not count as 'utterances' for purposes of clause D1.

2.4. Other properties of back-channel feedback

This section surveys some properties which are common to most back-channel feedback but which we chose not to regard as definitional; it thereby justifies why our definition includes some marginal cases.

Back-channel feedback often expresses attention, understanding, or agreement. However, this is not necessary (Rosenfeld and Hancks, 1980; Schegloff, 1982). On the one hand, back-channel feedback can express more than these things. In our corpus, back-channel feedback sometimes takes the form of words expressing degree of agreement, words of judgment, and words of sympathy and approval, among other things. It is generally agreed that such expressions, sometimes called 'assessments', pattern in much the same way as the simpler vocalizations, sometimes called 'continuers', and can be treated together (although there are some non-obvious differences, see Goodwin, 1986.) On the other hand, there are cases of back-channel feedback that do less. Not all signal attention; some signal boredom. Not all signal agreement; some signal skepticism. Not all signal understanding, often because there is nothing to understand, as in cases of disfluencies. Indeed, many back-channels seem to relate mostly to upcoming transmission of information, rather than to a previous transmission; they serve to indicate that the 'channel is open' or as an invitation to 'please continue'.

Back-channel feedback often consists of characteristic lexical items, such as *uh-huh* in English and *un* and *hai* in Japanese. For purposes of automatic discourse act classification, these specific lexical items are a valuable cue (Jurafsky et al., 1998). However, there is great variation in the words and phrases used in back-channel feedback, and infinitely many non-lexical vocalizations are possible (section 6.5). There are even cases of laughter, coughs, and sniffs which seem to function in the same way as does regular back-channel feedback. Thus, it does not seem wise to define back-channel feedback as a set of lexical items.

Most back-channel feedback is short. Other things being equal, longer utterances have a stronger tendency to interfere with the other speaker (Mueller, 1996), or at least require him to pay attention, violating D3. That back-channel feedback be short is sometimes taken as definitional (Koiso et al., 1995), but extreme shortness is not always required, provided that the contributions do not serve as interruptions. Long back-channel feedback often results in overlapping talk, and often seems to express enthusiasm. Incidentally, this seems more characteristic of female–female dyads, both in Japanese and English (Tannen, 1990).

Back-channel feedback is sometimes defined as those utterances which 'do not take the floor', and/or 'are not full turns'. The intuition behind these phrasings is captured in clause D3 of our definition, in so far as 'requiring acknowledgment by the other' is characteristic of full turns. Conversely, if the speaker carries on talking, then he is generally treating the listener's response as a back-channel. This is not a necessary condition, of course; the speaker is free to do what he wants. There are thus cases where the listener produces something probably intended as a back-channel, but the speaker then falls silent, or even responds explicitly to the back-channel.

Most back-channel feedback seems to appear while the other 'has the floor' or during the other's 'turn' or 'speakership'. However, these notions are too problematic to use in definitions. In particular, often these terms could not be applied to our data, in cases when both participants were talking simultaneously, and in cases where the speaker seems to have stopped and be waiting for the other to take a turn, but the respondent produced a 'perverse passive' back-channel as a coy way of refusing to take a turn (Jefferson, 1984). Incidentally, those working with the notion of floor often consider requests for clarification and answers to questions to be back-channel feedback (Duncan and Fiske, 1985; R. Hayashi, 1996).

Back-channel feedback is often characterized as serving to make the conversation go smoothly, but this is not useful as a criterion for deciding whether a specific utterance is back-channel feedback or not. On the respondent's side, it is generally impossible to tell what any given instance of back-channel feedback is intended to mean or do, let alone relate that to something as nebulous as smooth conversation. And on the speaker's side, it is not generally possible to tell what, if any, effect any single contribution has. That is, back-channel feedback usually seems to have no immediate dialog effect, and even longer-term effects, such as encouraging the speaker to keep talking, are highly variable (Siegman, 1976). Also, it is occasionally the case, in our corpus at least, that back-channels are produced which are almost certainly too quiet to be audible to the speaker.

Incidentally, we use the term 'back-channel feedback' because it is neutral with respect to discourse function. The terms 'assessment', 'acknowledgment', and 'reactive token' highlight the relation of these items to the previous utterance, whereas terms like 'continuer' highlight the relation of these items to upcoming utterances. But most instances seem to bear both functions: a backward-looking function and a forward-looking function. To avoid focusing on either function, we use the neutral term, back-channel feedback.

2.5. *Two comments on alternative definitions*

Thus we have extended the category of back-channel feedback from the clear cases to something more general. In doing so, we have disregarded some factors that sometimes seem important, and so our category cross-cuts the categories proposed by other researchers. We do not think this is a problem; different definitions are appropriate for different research aims. For example, it is interesting to consider the various functions of words like *okay* and *yeah* and *mm* in English, treating together their back-channel uses and their turn-opening uses (Jefferson, 1984; Novick and Sutton, 1994; Clancy et al., 1996; Gardner, 1997); the uses of items such as *oo* as both back-channels and fillers (Schiffrin, 1987); the back-channel, filler, and answer uses of Japanese *un*, *ee*, and *hai* (Kawamori et al., 1994), and the back-channel, turn-initial, and post completion uses of Japanese *un* (M. Hayashi, 1996).

This paper has discussed only back-channel feedback in casual conversation. In studies of task-oriented and business-like dialogs, the term 'back-channel' (and '*aizuchi*' in Japanese) is sometimes used loosely to refer to acknowledgments of var-

ious kinds, including responses to instructions, requests, suggestions, and new information (Boyle et al., 1994; Kawamori et al., 1994; Koiso et al., 1998; Okato et al., 1996; Okada, 1996; Araki et al., 1997). These resemble back-channel feedback in being short and often using the same lexical items, including *okay* in English and *hai* and *un* in Japanese. But there is a major difference, namely that most of these seem to be expected by the speaker, who waits for them to appear before continuing, violating clauses D2 and D3 of our definition.

3. The low pitch cue

Armed with this definition, we labeled occurrences of back-channels in conversation corpora, and then sought for prosodic features of speaker's utterances which reliably preceded such listener behavior. (The details of the methodology are given below, section 6.8.)

We have found that, in American English conversations and in Japanese conversations, after the speaker produces a region of low pitch lasting 110 milliseconds the listener tends to produce back-channel feedback (Ward, 1996; Ward, 1997b). (This can be seen in figures Figs. 1 and 2.)

This correlation is not something that speakers are consciously aware of. However, after it has been pointed out, it is apparent to the unaided ear. In particular, (1) as an eavesdropper, you can observe it, (2) as a listener, you can use it to determine when to produce feedback, and (3) as a speaker, you can use it to deliberately elicit feedback. In each case, the connection clearly works most of the time.

The correlation is not perfect, as discussed below, but it does seem that low pitch regions 'indicate places where back-channel feedback is especially appropriate'. This being an unwieldy expression, we will say that a low pitch region is a 'cue' for back-channel feedback.

3.1. Precursor findings

Our finding has several precursors. It is refinement of Sugito's observation that a low pitch point seems to correlate with back-channels in Japanese (Sugito, 1994). It also probably accounts for the finding that back-channel feedback is especially welcome at junctures between phonemic clauses (Dittmann and Llewellyn, 1967), and at "the ends of intonation units that have non-final intonation contours" (Clancy et al., 1996), to the extent that the low pitch region is one of the acoustic features that serves, perceptually, to mark or foreshadow these. It also relates to the general fact that the ends of utterances, clauses, and similar things tend to be marked with low pitch, in the form of declination or boundary tones, and the fact that hearers are logically more likely to produce back-channel feedback after such units end. What is novel is the specification of exactly what sort of low pitch correlates with back-channel feedback, namely a *region* of low pitch.

Our description of the cue is also compatible with, although more general than, Noguchi's recent suggestion that those back-channels in Japanese which occur dur-

ing the speaker's pauses are cued by utterances which end in a low pitch and have a final region of flattish pitch (Noguchi et al., 1998).

3.2. Quantitative statement

In order to measure the strength of this cue, as compared with other possible cues, we formulate the following predictive rule for English:

Upon detection of
(P1) a region of pitch less than the 26th-percentile pitch level and
(P2) continuing for at least 110 milliseconds,
(P3) coming after at least 700 milliseconds of speech,
(P4) providing you have not output back-channel feedback within the preceding 800 milliseconds,
(P5) after 700 milliseconds wait,
you should produce back-channel feedback.

The exact parameters of these rules were chosen to maximize correspondence to corpus data, as discussed in section 4. For Japanese some parameters are different: P1 = 28, P2 = 110, P3 = 700, P4 = 1000, and P5 = 350. The differences between the English and Japanese versions of the rule are discussed in section 6.3.

Conditions P1, P2, and P5 express the core of the rule. Condition P3 reflects the intuition that a listener should not produce a back-channel before the speaker has got started, and P4 the intuition that back-channels should be appropriately spaced. P3 and P4 are discussed further below.

An example of such a cue appears near 39,000 in Fig. 1, at the word *gotten*, with the resulting prediction seen as the wide oval near 39,600. In Fig. 2 there is a cue near 48,000, at the word *ne*.

3.3. Communicative functions of low pitch regions

This subsection considers the functions of regions of low pitch, and discusses some typical contexts where they function as low pitch cues and cases where they do not.

First, low pitch regions often occur at points where the speaker considers that he has transmitted some information. At these points, it is logically appropriate for the hearer to confirm receipt or understanding or interest with a back-channel. We can think of these low pitch regions as conveying 'this completes that thought, did you follow?' Sometimes what has been transmitted is a complete new fact or proposition, but not always – 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 sometimes appears before the speaker has completed a grammatical phrase or full proposition, and may even take the form of completing the speaker's thought or sentence.

Such low pitch regions often co-occur with completion of a grammatical clause. Concomitantly, they also often co-occur with related lexical items, especially in

Japanese, where they frequently appear along with clause connectives, most often *kara* 'because', *-te* 'and' and *kedo* 'but'. Note that these clause-ending markers do not function as sentence ends, at least not prescriptively, but rather indicate some degree of incompleteness (Mizutani, 1988).

Some of these cases of completion are, perceptually, 'turn' ends, and in such cases, the low pitch rule generally gives rise to incorrect back-channel predictions. In both languages, abrupt drop in pitch and/or energy often seems to indicate end of turn (Noguchi et al., 1998; Koiso et al., 1998).

Rather less often, a low pitch region which seems to mark the conveying of information appears with a repetition of a previous word, produced for emphasis or clarity and/or when recovering from a false start, especially in English. In such cases also, it often welcomes back-channel feedback; we can perhaps consider it to convey 'I said it again, did you get it that time?'

Second, low pitch regions also occur frequently with disfluencies and markers of formulation difficulties, especially in English (Ward, 1999). In these cases we can think of the low pitch region as saying, 'I'm stuck, but keep listening, something meaningful will come out soon'. In English the single most frequent lexical item appearing in low pitch regions is *the*, often in the lengthened, unreduced pronunciation indicating formulation difficulty (Fox Tree and Clark, 1997); the next frequent two are *and* (usually lengthened), and *um*. In Japanese, these regions also appear fairly often with disfluency markers, such as *nanka*. A related communicative function is taking the floor before actually saying anything, where the speaker utters some kind of call for attention, or filler, and low pitch regions occasionally occur at these times.

In both languages, the disfluency with low pitch sometimes elicits back-channel feedback, presumably functioning as encouragement to continue. Many such disfluencies, however, do not seem to welcome back-channel feedback, and this gives rise to predictions which are incorrect. However, clause P3 of the rule serves to prevent many of these, as disfluencies tend to occur early in an utterance. Interestingly, Shriberg et al. have shown that the discrimination between disfluent and fluent speech can be detected fairly well using prosody alone (Shriberg et al., 1997).

Third, low pitch regions in Japanese often occur with sentence-final particles, especially *ne* 'you know', which typically serves to 'seek agreement' or 'invite collaboration' (Cook, 1992) and is known to be associated with several turn-taking operations (Tanaka, in press), and is the word most frequently appearing with low pitch regions.

Fourth, a low pitch region often occurs together with back-channel feedback itself. In the corpora such cases occasionally elicit a confirmatory uttering or sigh; our definition of back-channel feedback however excludes such responses, that is, responses to back-channels are not themselves considered to be back-channels (section 2.3). Thus, low pitch regions on back-channels occasionally result in incorrect predictions, although clause P3 of the rule prevents most such problems, since back-channel feedback usually is short. Interestingly, Jurafsky et al., have suggested that the discrimination between back-channels and other dialog acts can be done to some extent by using prosody alone (Jurafsky et al., 1998).

Fifth, there are cases where 110ms of low pitch occurs as part of a substantially longer region of low pitch, with a special meaning. These cases include groans (in Japanese at least), and utterances with reduced pitch range, as in parentheticals (in English at least) and 'self-directed speech', that is comments said 'under the breath'. Although rare, these instances generally cause the rule to make incorrect back-channel predictions.

3.4. More about the parameters of the rule

P1 refers to the maximum pitch; a region counts as a 26th-percentile pitch region if all pitches in that region are less than the 26th-percentile value. Perhaps P1 should be defined instead in terms of the average pitch, or perhaps a weighted average pitch, depending on the details of how people perceive pitch levels.

P3 is a heuristic approximation to the intuition that you shouldn't produce a back-channel unless the other person has said something. In practice, it prevents predictions in response to utterance-initial low pitch regions, which are typically back-channels, floor grabbing tokens, fillers, and other disfluencies (section 3.3). While P3 works well most of the time, a rigid 700ms cutoff is obviously too simplistic, not only theoretically but also practically. On the one hand, this sometimes incorrectly prevents predictions in response to short but content-rich utterances. On the other hand, this fails to prevent predictions of back-channels as responses to the occasional very long back-channel.

P4 encodes the intuition that it is generally appropriate to leave some decent interval between back-channels; were it not for P4, the rule would sometimes predict several back-channels in the span of a second. P4 is, however, quite unreliable. It is a source of missed predictions (cases where in the corpus two back-channels occur in quick succession) and of inaccurate predictions (cases where it seems that the reason why a respondent passed up the opportunity to produce a back-channel was that she had fairly recently produced one). Another problem arose because the task definition (section 4.2) required P4 to be computed, not with reference to the track of the corpus that was being emulated, but with reference to the rule's own previous predictions. This caused propagation of errors, where the rule first generated an incorrect prediction, and P4 therefore suppressed a possible later prediction which would have been correct.

Incidentally, in preliminary work we set the timing (P5) relative to the end of the low pitch region rather than to the point of detection of a low pitch region. The current definition gives slightly better performance and is somewhat more psychologically plausible, in that it allows earlier detection of cues, which makes it easier to understand how humans can respond as quickly as they do.

The current rule combines the conditions with simple conjunction – all must hold true for a prediction to be made. While better prediction quality might be obtained by using a 'fuzzy' or more sophisticated combination of the various factors, the current formulation does have the advantage of simplicity.

It turns out that the exact parameter settings which give best performance vary among speakers, reflecting, at least in some cases, individual differences in conversational style.

3.5. The details of the computation

The values for parameters P1 through P5 are not eternal constants. Among other things, they depend on the details of how pitch and energy are computed. This is why the values of the parameters of the current rule differ from those reported in preliminary work (Ward, 1996) – all were re-tuned to optimize performance after changes to the details of the implementation. This subsection describes the key points of the current implementation.

The speech signal is processed at 8000 samples per second, 8-bits per sample, μ-law format (Owens, 1993). For the corpus-based evaluation, this was obtained by down-sampling from DAT format; for the live experiments (section 6.4), this is done directly using the built-in microphone jack and analog-to-digital converter of a computer. The pitch and energy are computed for every 10ms frame.

Regarding P3, a measure of energy is computed for each frame and a histogram of energy values is made. The average of the first percentile value and the 99th percentile value is the threshold; frames with energy level greater than this are considered to be speech. (While there are better ways to distinguish speech from non-speech, this was generally adequate for our data.) For grouping speech frames into speech regions, gaps of up to 250ms of non-speech are allowed. This works fairly well, except for geminate consonants in emphatic speech.

Regarding P1 and P2, pitch is computed using Hirose's pitch detector (Hirose et al., 1992). The low pitch threshold is set at the 26th percentile pitch level; thus frames with a pitch lower than this are considered to be low pitch frames. Pitch values 'spread' to subsequent frames at which no pitch was detected, up to 80ms later, provided that all intervening frames are speech frames (not silence). This serves to spread pitch across consonants, and, more importantly, compensate for some failures of the pitch tracker, particularly in regions of vocal fry (creak) and when the volume trails off at the end of a phrase. The condition that the intervening frames contain speech prevents spreading before a pause, which matters because a short region of low pitch before a pause is often an indication of finality, or 'turn end'.

To determine the energy and pitch thresholds, the system computes the distributions of energy and pitch values. These distributions are computed over the input up to the current point; that is, they are computed on-line, with no lookahead. The system would probably get slightly better performance if it computed the energy distribution over the entire conversation, but we wanted our implementation to work in the same way for the live experiments. For the pitch distribution, only pitch values in the past 50 seconds are considered, which makes the value of the 26th-percentile pitch sensitive to long-term changes in pitch range; this is useful as adaptation to local increases in pitch range during interesting portions of the conversation.

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

4. Corpus-based evaluation

Our hypothesis is that the rules of the previous section are indeed rules of English and Japanese conversation. This section measures how well these rules perform, compared to some other possible rules, by computing the degree of match between the predictions made by the rules and the actual occurrences of back-channel feedback in a corpus of conversations.

4.1. The corpus

The point of the corpus being mostly to have some data to use for judging rule performance, we did not attempt to make the corpus representative, nor uniform, nor balanced. Specifically, we did not control for conversants' sex, native dialect, education level, relative status, motivation, or mutual familiarity, nor for conversation location, topic, time into the conversation, speech rate, and so on, although some of these factors will undoubtedly play role in a complete account of back-channel behavior.

The size of the corpus was determined by the desire to report statistically significant results in section 4.4, and by a feeling that, after a certain point, the process of gathering and listening to more data was not yielding significant new insights.

Each conversation had two participants. In most of the conversations, the participants were seated in such a way as to prevent them from seeing each other. Thus these conversations were similar to telephone talk. This of course meant that non-verbal back-channel cues, such as gaze (Rosenfeld and Hancks, 1980; Duncan and Fiske, 1985) were prevented, as were non-verbal forms of back-channel feedback, such as nods and smiles (Maynard, 1989). As a result, it is possible that our results do not extend to face-to-face interaction. In most cases the conversants were told of our interest in back-channels, but this did not appear to affect their behavior. Recording was done using head-mounted microphones in stereo onto DAT tape and the conversations were uploaded to a computer for labeling and analysis.

Our English corpus is 68 minutes total, consisting of 8 conversations, involving 12 different speakers (one speaker, the first author, participated in 5 of the conversations). All participants were native American speakers except one, who had lived in the U.S. since her early teens. Two participants were female, 10 male. The conversations were recorded in various locations in the U.S., except one recorded in Tokyo with short-term visitors. None of the participants appeared to be using either of the geographically or ethnically identified conversation styles that reportedly differ markedly from mainstream American practice (Erickson, 1979; Tannen, 1990). The five most frequent kinds of back-channel feedback were *yeah*, *uh-huh*, *hm*, *right* and *okay*, which is comparable with what is found in other corpora (Jurafsky et al., 1998). By the definition given in section 2, the English corpus has 359 back-channels.

Our Japanese corpus is 80 minutes total, consisting of 18 conversations, involving 24 different speakers, including the second author in one conversation. All samples were recorded in Tokyo, and all conversants were native speakers of Japanese in their twenties. Nine participants were female, 15 male. The five most frequent back-

channels were *un*, *ee*, *aa*, laughter, and *hai*; this is roughly in line with what has been observed in other studies of casual Japanese conversation (Maynard, 1989; Clancy et al., 1996). *Un* (actually usually pronounced as a nasalized schwa) predominates, accounting, with its variants, for almost a third of the back-channels. By the definition of section 2 the Japanese corpus has 873 back-channels.

4.2. The task

The task we set for our rule was, given any track of a conversation from the corpus,[1] predict where back-channels occur in the other track.

Specifically, our rule is invoked to make a decision every 10 milliseconds. That is, every 10 milliseconds it either predicts an instance of back-channel feedback or it does not.[2] (Note that, as explained below, the domain of validity of a prediction is 1000 milliseconds, and, as seen in section 3.2, P4 means that the rule always leaves a space of at least 800 milliseconds between predictions.)

A well-defined task such as this has the merit of providing a simple evaluation technique, adequate as a way to compare the performance of alternative proposed rules. We needed this to discover which parameter values gave the best results for our rule, and to compare the performance of our rule to other rules. This technique does not, however, provide an absolute measure of predictive power, since various problems mean that no rule could achieve perfect performance on this task (see section 5.5).

Specifically, a better rule is one that has higher 'coverage', that is, the number of back-channels which the rule manages to predict, and higher 'accuracy', that is, the fraction of predictions which are correct. In many situations, however, two rules rank differently on the coverage metric and the accuracy metric. That is, there is a 'trade-off' between coverage and accuracy. For example, for our rule, as the definition of 'low pitch cue' is made looser (with a higher value for P1 or lower values for P2, P3, or P4), coverage increases at the expense of accuracy, since more back-channels will be correctly predicted, but more incorrect predictions will also occur. Conversely, if the definition is stricter, accuracy will be higher but coverage lower. In such situations we prefer the rule which gives the best value for the *product* of accuracy and coverage. This figure of merit favors rules which attain good scores on both metrics, rather than excellent performance on one at the expense of the other.

---

[1] For all tracks, that is to say, for all 12 English speakers and all 24 Japanese speakers; this contrasts with research which limits itself to the modeling of one or a few speakers (Okato et al., 1996; Noguchi et al., 1998).

[2] Thus our task differs from that adopted by research which merely seeks for rules which discriminate between utterence-ends which are followed by back-channels and those which are not (Noguchi et al., 1998), which is an easier task, in that it exploits foreknowledge of whether a short pause will turn into significant one, and/or whether it will be followed by a change of speakership (Noguchi et al., 1998). Moreover, studies of the relationship between utterence-end prosody and back-channels inevitably fail to account for back-channels which do not follow pauses, overlapping with the speaker's utterence, and such cases are common.

There is another factor involved in performance: the frequency of prediction. It seems that, independent of where back-channels are produced, how often they are produced is also important. Optimizing prediction frequency is not very interesting – it is trivial to predict at the same frequency as in the corpus by adjusting the parameters of a rule. Rather, any high-coverage, high-accuracy rule is acceptable as long as it does not produce vastly more or vastly fewer back-channels than people do. Given the wide variation in back-channeling frequency, we chose here to target the frequency of the more supportive, friendly listeners in the corpus, who tended to produce back-channels more often than average.

Although it is generally good practice to evaluate a classification algorithm using a previously unseen corpus of test data, we do not do so here, because our rule, having only five free parameters, stands in little danger of being inadvertently tailored to one specific set of training examples.[3]

4.3. Our measure of successful prediction

We wanted to count a prediction of back-channel feedback as correct if it was at 'essentially the same time' as actual back-channel feedback in the corpus. Again, a perfect measure was not required, merely one good enough to let us compare the performance of different rules. We therefore operationalized this by counting a prediction as correct if the predicted onset of a back-channel was within 500 milliseconds of the onset of an observed back-channel.

The decision to tolerate misalignments of up to 500 milliseconds was based on informal judgments of 'how much earlier or later a back-channel could appear and still sound appropriate' in various contexts. We found that it was not uncommon for back-channels that were timeshifted as much as 500 milliseconds to still sound fairly natural. This result differs from what can be suggested by passively listening to a conversation, where it often seems that each back-channel is produced at a precisely appropriate time, but that impression is probably misleading (O'Connell et al., 1990). Of course, a simple 500ms window is not completely satisfactory. First, one could use a narrower window, to ensure that anything within that window would be quite likely to be natural (Okato et al., 1996). Second, one could try to compute degrees of naturalness, as a function of the degree of misalignment. Third, one could judge each prediction individually, in context, by hand. We opted for generosity, simplicity, and convenience, as we didn't need an absolute standard, but only a way to compare the performance of alternative proposed rules.

Note that using the ±500ms window for matching means that, since we are using a 350ms value for P5 in the Japanese rule, our evaluation will give the rule credit for predicting back-channels whose onset precedes the detection of the low pitch region. While this is reasonable for judging whether it would be acceptable to produce back-channels using the rule, it is of course inappropriate for judging whether the

[3] This would not be true had we used a decision-tree learning algorithm, as for example used by Koiso et. al. (1998), which can give an arbitrarily high predictive score on the training data.

rule could account for human behavior; thus, below we also report results with a tighter window.

There was an additional proviso: each prediction had to match a different observation. Thus, if the rule predicted back-channels at 1200 milliseconds and at 1800 milliseconds into the conversation, and a back-channel actually occurred at 1500 milliseconds, that would count as one accurate prediction and one incorrect prediction. This proviso was problematic in cases where a long back-channel covered the time span between two predictions, but such cases were rare.

We decided to use onset as the reference point, because in most cases the onset of a back-channel seems to be its most salient point. Of course, given our wide (±500 milliseconds) window, this choice was not usually critical, since most back-channels last only a few hundred milliseconds.

4.4. Results

Performances of our rules for English and Japanese are seen in Tables 1 and 2. For comparison, the tables also show results for prediction rules which do not use low pitch regions, specifically, making predictions at random while obeying P3, P4, and P5. Note that, on a first approximation, varying the frequency of random prediction does not affect accuracy.

Table 1
Performance of various rules for predicting back-channel feedback (English)

Predictions from	Coverage	Accuracy	Figure of merit
Low pitch regions (section 3.2)	48% (172/359)	18% (172/936)	.088
Random (section 4.4)	22% (80/359)	13% (80/618)	.029
Utterance end (section 5.1)	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 (Japanese)

Predictions from	Coverage	Accuracy	Figure of merit
Low pitch regions (section 3.2)	56% (496/873)	34% (496/1447)	.195
Random (section 4.4)	25% (222/873)	24% (222/915)	.062
Utterance end (section 5.1)	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
After *ne* (section 5.1)	31% (273/873)	23% (273/1191)	.072
After *kedo* (section 5.1)	21% (181/873)	26% (181/691)	.054
After *kara* (section 5.1)	14% (122/873)	23% (122/533)	.032
Eavesdropping human judge	~95%	~61%	~.58

The low pitch based rules do better than random: for example, the accuracy was 18% versus 13% for English and 34% versus 24% for Japanese. This is true both on average, as seen in the tables, and for most speakers in most conversations: specifically, for English in 15 cases out of 16 (2 sides times 8 conversations) the figure of merit was higher for the low pitch rule, and for Japanese it was better for 34 out of 36 cases.

The rule was superior to random prediction also when judged using tighter values for the permitted misalignment window, ranging from 100 milliseconds to 400 milliseconds. Also, adding the constraint that predictions could not match corpus back-channels more than 350ms earlier than the prediction, in the Japanese case, to count only those cases where the rule might account for human behavior (section 4.3), this gave a coverage of 49% and an accuracy of 29%, still higher than chance.

5. Other factors

As seen by the coverage and accuracy results above, the low pitch cue accounts for many of the back-channel occurrences, but not all. This section surveys some other factors involved in the production of low pitch regions and in the cueing of back-channel feedback.

5.1. Correlated cuing factors

A factor which seems likely to account for many back-channels is end of utterance or pause. Certainly there are times when awkward silences lead to the production of a back-channel, and such occasions are quite salient. However, if the notion of end of utterance is straightforwardly taken to mean the onset of silence, without regard to prosody, then this is not statistically a good cue for back-channel feedback. Specifically, the accuracy of end-of-utterance-plus-low-pitch predictions is not significantly better than the predictions from low pitch alone (19% versus 18% for English and 32% versus 34% for Japanese), and end-of-utterance-plus-no-low-pitch is less accurate than random prediction (5% versus 13% for English and 16% versus 24% for Japanese), as the tables show. (The results are for the best silence-based rule we found, namely one which predicts a back-channel in response to 150ms of silence, subject to clauses P3, P4, and P5 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, and indeed such cases are common in the corpus (one is seen near 39,300 in Fig. 1).

Rising intonations, including 'uptalk' (also known as 'high-rise questions', see Hirshberg and Ward, 1995) also sometimes elicit feedback. These seem generally to indicate insistence: in demanding a response, in demanding agreement, in giving instructions, or in making a suggestion. Indeed, it has been claimed that in Japanese, strongly rising intonations *always* elicit a response (Saito, 1997). (Because of this, some instances of an utterance in response to such pitch patterns are to be considered to be full responses, not back-channel feedback, according to our definition.) In

some cases where uptalk does prefigure back-channel feedback, the rise is preceded by a region of low pitch; it often seems that a rising intonation at the end merely adds a nuance of insistence to the basic low pitch cue.

Vowel lengthening often precedes back-channel feedback, in both English and Japanese (Maynard, 1989), especially in cases of disfluencies and agreement-seeking particles. This generally seems to appear together with the low pitch region. Indeed, lengthening can perhaps be analyzed as 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'. This hypothesis is supported by the fact that lengthening seems to occur less often when the low pitch region falls on longer words and phrases, such as *da yo ne* 'COPULA I'm-telling-you you-know'.

Sugito (1997) has observed that speaker disfluencies also frequently precede back-channels in Japanese. As noted above, this often appears together with low pitch regions.

For Japanese, Mizutani (1984) has hypothesized that low volume in an utterance cues back-channel feedback. In our data this did not seem to be an important factor.

In many cases, back-channel feedback is preceded by a region of vocal fry (creak or creaky voice). Most such cases meet the criteria of our low pitch rule, and so it is likely that vocal fry is not an independent cue.

In other cases, back-channel feedback is preceded by certain lexical items, including *ne* and other particles, clause connectives, and disfluency markers. The first two categories seem to be plausible candidates for cues (Mizutani, 1988; Maynard, 1989). In our corpus, however, they do not turn out to be strong cues; the accuracy of back-channel predictions based on such words is near chance (*ne* 23%, *kedo* 24%, *kara* 23% versus random 24%), as seen in Table 2. Moreover, most of the time when these words appear prior to back-channel feedback, they appear together with a region of low pitch.

It has been found that completion of a 'grammatical clause' is also a cue for back-channels (Duncan and Fiske, 1985; Maynard, 1989; Clancy et al., 1996), despite the difficulty of identifying grammatical clauses in unscripted conversation (Ford et al., 1996). Grammatical completion also correlates with low pitch regions.

Incidentally, there appear to be significant differences in the factors that cue non-backchannel acknowledgments (section 2.5); in particular, their occurrence seems to depend more on meaning (Kawamori et al., 1994), on end-of-utterance silence (Fukuchi, 1997), and on prosodic features other than low pitch (Koiso et al., 1998), than does the occurrence of back-channel feedback.

5.2. *Independent cuing factors*

Speakers sometimes enunciate a word, pronouncing it carefully, slowly, and in a fairly higher pitch, when introducing an unfamiliar referent into the discourse. This often seems to elicit back-channel feedback, often an echo of that word.

A rare form of back-channel feedback is a vocalization as one speaker yields the floor, after an occasion when both speakers inadvertently begin talking at the same time.

5.3. Counter-indicating factors

There are also some 'counterindicating factors', which seem to serve to override or suppress responses to the low pitch cue.

One is when a speaker abruptly starts a new utterance at a high pitch shortly after producing a low pitch region, as if changing his mind about whether he wants back-channel feedback. In most, but not all, such situations the respondent in the corpus did not produce a back-channel, and in those cases the rule led to an incorrect prediction.

Another is when the low pitch region occurs as part of some larger pitch contour. For example, an upturn in pitch at the end of a region of low pitch sometimes seems to turn a cue for back-channel feedback into a cue for a substantive comment or answer, at least in Japanese. Perceptually, such contours seem to signal the end of a 'turn' or the completion of a question, challenge, or suggestion that required a reply. Predictions of back-channels at places where the speaker in the corpus took the floor (and listening suggested that a back-channel was truly inappropriate), were fairly common, accounting for 17% of the total number of incorrect predictions.

Third, more speculatively, the type of conversation activity may play a role. In narrative and explanation, which accounted for most of our corpora, the low pitch region is mostly a back-channel cue, but in other activities, such as banter, question and answer, instruction, suggestion giving, attempts to find something to talk about, musings, and ritual greetings, it sometimes seems to invite a full turn or no response.

5.4. Interactions among factors

For most of these factors, we cannot yet say whether they are independent cues to back-channel feedback, whether they precede back-channel feedback only to the extent that they co-occur with low pitch regions, or indeed whether the low pitch regions precede back-channels only to the extent that low pitch occurs with these factors. Because cues of several types so often appear together redundantly, quantitative analysis of the effectiveness of combined cues, and of conflicting cues (Ford and Thompson, 1996; Furo, 1997; Koiso et al., 1998), is clearly required as a next step. Another important question, for the sake of identifying the cues that listeners actually respond to, is whether some cues are typically evident slightly earlier than others.

5.5. How much is not yet explained

To summarize the above discussion, it seems that no other single factor can account for all the occurrences of back-channels that low pitch regions can. This even seems to be true for meaning: it seems that no simple notion of meaning can account for all the occurrences of back-channels that low pitch regions can (sections 3.3 and 2.4).

However, it also seems that there are many factors which may account for occurrences of back-channels, including some occurrences which the low pitch rule does not explain. This subsection estimates how much of the data is in fact not explained by the low pitch rule. However, this is not trivial, due to two shortcomings of our evaluation method (section 4.2).

The first problem arises from the presence in the corpus of many cases where both conversants are talking at once. In such cases, there was not really any option of producing back-channel feedback, but our rule, only allowed to use the information in one track of the conversation, could not know this.

The second problem is that the actual pattern of back-channels is not necessarily the only possible pattern; in particular, there are differences among speakers in back-channel production style, most obviously in frequency. A related problem is that a rule can predict opportunities, but respondents do not choose to produce back-channel feedback at every opportunity. The most common alternative, always available, is to be silent. Another alternative is to produce a full turn, especially a request for repetition or clarification.

To attempt to measure the significance of these problems, we examined, for the Japanese corpus, all cases where the rule made an prediction which did not correspond to a back-channel in the corpus. Of these, 16% coincided with an utterance in progress, and thus accurate prediction could not have been expected, because of the first problem above. Another 44% of the incorrect predictions were cases where a back-channel could naturally have appeared, as judged by the second author, but in the corpus there was silence or, more rarely, the start of a turn. That is, these incorrect predictions seem to be due to the second problem above, inter-speaker differences in back-channel behavior. This interpretation is supported by the finding that, if the rule is judged as a model of a single speaker (in this case, the speaker identified as JH), its performance is much better: a coverage of 69% (57/83) and an accuracy of 66% (57/86).

On the other hand, there were cases where a back-channel was present but the rule failed to make any prediction. While most of these seemed to be true failures of our rule, a few were not, due instead to what appeared to be idiosyncratic behavior by some speaker – places where the second author listening to the corpus felt that a back-channel was strange, and that most Japanese speakers would never produce a back-channel in such circumstances. However, these were rare. Another uninformative cause of prediction failures was the fallibility of the pitch tracker used.

Working from these estimates, it seems that the greatest possible coverage one could hope from any rule, measured using this method, would be about 95%, and the greatest accuracy about 61%, as seen in the last row of Table 2. We can also sum up the accuracy of the Japanese rule as follows: 34% of its predictions were accurate, about 27% were inaccurate, and the remaining 39% were probably not wrong.

Thus, there is a lot that the present rule does not account for. Some of the unexplained data could probably be accounted for in terms of other cues from the speaker, including the factors discussed above. A full account would, however, doubtless require more consideration of the respondent's role; since there are cases

where the occurrence of back-channels depends not only on the speaker's cues, but also on respondent-internal factors, such as his speed of understanding and his attitude to the speaker, among a multitude of other factors.

5.6. Failure analysis

Much of the discussion above was based on analysis of cases where our low pitch rule failed. It might seem appropriate to also report the raw statistics on the frequency of each cause of failure. We do not do so because failures are so hard to analyze conclusively. For example, many more occurrences of back-channels could have been predicted, had one or another of the parameters of our rule been a little more lax. However, looking at the big picture, such increases in coverage would be at the expense of accuracy, as we found by systematically varying each parameter and re-running the variant rules over the entire corpus. Another reason why failures are hard to analyze is that many occurrences of back-channel can plausibly be accounted for by several factors, with none clearly decisive.

6. Open issues and speculations

This section offers some observations and speculations on various open issues.

6.1. Relation to other uses of pitch

One alternative approach to accounting for the facts we have described is to argue that 'low pitch regions are mere epiphenomena, resulting from the lack of 'stresses' or 'accents' and the accompanying high pitches, in regions of speech which convey no new or important information – and that it is the lack of new information, or its prosodic correlates, which are the true cues which encourage the listener to back-channel'. We consider this unlikely, given that low pitch regions do appear to bear specific communicative functions (section 3.3). However, the question of how the low pitch cue actually does relate to lexical pitch accent and sentence-level prosody is completely open. This is a daunting issue, not least because it is not clear to what extent the insights arrived at by studies of read speech, monolog, controlled dialog, and so on, carry over to unscripted, casual conversation (Schuetze-Coburn et al., 1994).

The existence of 110ms low pitch regions as meaningful prosodic features may also bear on an issue in the theory of intonation, namely, the question of whether it is possible to devise an abstract level of representation that mediates between the language level and the sound level (Ladd, 1996), or whether an account of the uses of prosody must directly refer to acoustic features, as our finding may suggest. Alternatively, if these low pitch regions are relegated to the domain of paralinguistic phenomena, it is the question of the relation between the linguistic and paralinguistic aspects of prosody that takes on a renewed significance.

6.2. More precise timing

The current rule certainly does not account for the exact timing of back-channel feedback. At least three additional types of factors need investigation. First is the question of what other factors, presumably mostly prosodic, the listener uses to determine exactly when to produce back-channel feedback. Second is the question of what shades of meaning the respondent can convey by producing back-channel feedback a few hundred milliseconds earlier or later (Ward, 1998). Third is the question of the role of human processing time requirements in the timing of back-channel feedback (Ward, 1997a).

6.3. Differences between English and Japanese

Comparisons between the two languages are difficult because, among other problems, our English conversants were generally slightly older, mutually less familiar, and talking more seriously, than our Japanese conversants. Thus we cannot say, for example, whether it is significant that the most predictive value for low pitch region length (parameter P2, 110 milliseconds) is the same for both languages.

However, there are three differences between Japanese and English that may be meaningful. First, the accuracy of our English rule is lower. It seems that English speakers tend to take up fewer of the opportunities to produce back-channel feedback. This result is compatible with the results of comparative studies of the frequency of back-channel feedback in Japanese and English (Maynard, 1989; Clancy et al., 1996). This difference has been convincingly related to various cultural differences in communication style (Mizutani, 1984, 1988; LoCastro, 1987; Maynard, 1989; White, 1989; Yamada, 1992; Strauss and Kawanishi, 1996; Clancy et al., 1996; R. Hayashi, 1996; Maynard, 1997), and to grammatical differences between the two languages (White, 1989; Fox et al., 1996). Second, the coverage is lower. One factor that may be involved is the presence of more uptalk in English. Another may be that English speakers tend to use grammatical completion relatively more as a factor in deciding when to produce back-channels (Maynard, 1989; Clancy et al., 1996). Yet another factor is that English respondents may be responding less automatically, not simply following cues from the speaker. Third, English respondents let more time lapse between the point when the cue is heard and the point where the back-channel is produced (P5). It is interesting to speculate that some of the huge perceived differences in conversation style and culture may reside in this 350ms timing difference.

6.4. An automated listener

Wanting to determine whether producing back-channel feedback in accordance with the low pitch rule meets the expectations of the conversation partner, we ran experiments where an unsuspecting subject was induced to 'converse' with a system incorporating the low pitch rule. The system produced a back-channel whenever the subject produced a region of low pitch. In general, third party judges listening to the

conversations could distinguish the low pitch based back-channels from the randomly produced ones; the former sounded natural and the latter sounded odd, with clear cases of inappropriate back-channels and of inappropriate silences when a back-channel was called for. Those who were actually talking to the system, however, were not obviously affected by the timing of back-channels, as Siegman (1976) also found. Further discussion of experience with this system and its implications appears elsewhere (Ward and Tsukahara, 1999).

6.5. *Sound and meaning in back-channel vocalizations*

The fascinating question of the content of back-channel feedback – what words are used in what situations with what meaning – is beyond the scope of the paper. However, we cannot resist mentioning that, for the many back-channels which are non-words, or 'grunts', the most parsimonious analysis may be that each of these vocalizations is custom-made for the occasion, with each of the various elements of the pronunciation of vocalization contributing some element of meaning (Takubo, 1994; Ward, 1998); that is, sound symbolism may be present.

6.6. *The psychological status of the low pitch cue*

So far we have shown that low pitch regions can account for some back-channel behavior. The question of whether there is actually a causal relation – whether these actually are cues – is ultimately a psychological one, probably decidable only by controlled experiment. Tentatively, however, it does seem possible that speakers sometimes may produce back-channels in direct response to the low pitch cue. In daily life, it is sometimes possible to produce appropriate back-channel feedback without paying attention (Dhorne, 1980), and do so well enough to get away with it, for at least 5 or 10 seconds, as when reading a newspaper while someone is talking to you.

Thus we might speculate, depending on the theoretical perspective, that respondents sometimes produce back-channels as a 'reflex' response to low pitch regions; or that the low pitch cue is processed in a separate 'channel', different from the channel used for understanding words and meaning; or that detecting back-channel cues and responding to them involves a mental 'module' distinct from that involved in the uptake and conveying of content in spoken language; or that back-channel cues and responses are part of a special 'modality'; or that they constitute another 'dimension of interaction'. (We certainly do not mean to suggest that all back-channels are produced in this way, without reference to words or meaning. Nor do we mean that there is a 100% correlation, with no effects of other factors. Also, while an account in terms of a direct response can explain in part whether a back-channel will occur, it may have nothing to say about which back-channel is to be produced.)

The idea that back-channeling sometimes involves a direct response is logically plausible, for two reasons. First, assuming a participant in a conversation generally has to go through three stages of processing to contribute – namely comprehension of the other's words, choice of response, and utterance – then there can be an advantage to being able to omit, or radically simplify, the first two stages, in that this lets

the respondent economize on mental effort. Second, to be a good listener requires, minimally, two things: being responsive to the other speaker, and not interrupting him. There is a trade-off between these two goals, since the more one speaks, the more likely one is to disrupt the other's utterances. To make the right decision, and make it quickly enough to be able to produce feedback at an appropriate time, is hard. Relying on a direct response to a prosodic cue may be the only way to respond fast enough.

6.7. What's interesting about conversation?

Conversations often seem to have a life of their own, for example when arguments continue on beyond the point of being useful to either party. Introspectively also, it sometimes seems that one can get 'caught up in' or 'carried away with' a conversation. More analytically, conversations seem to involve something special, something present above and beyond the goals and actions of the two participants. Theoretical constructs which seem to capture some of the specialness of conversation include 'interactional achievement' (Schegloff, 1982), 'conversation rule' (Duncan and Fiske, 1985), '*kyowa*' (Mizutani, 1993), 'co-construction' (Maynard, 1989), 'joint action', 'joint strategy' (Clark, 1996), and 'conversation as collaboration' (Clark, 1996).

Underlying this feeling of 'something special' is, we speculate, the 'reflex' aspects of some conversation behaviors, that is, the way they involve direct links between stimulus and response that apply automatically and unconsciously. Of course, this property is not unique to conversation phenomena; it is clear that most of language behavior is too fast to introspect on, or deliberate about, before doing.

What is interesting about the low pitch cue is that the trigger for rule application is not internal to the speaker, but an external stimulus. Thus, our viewpoint entails the idea that a respondent in a conversation is to a large extent controlled by the other participant, which may seem a strange perspective to some, as it sits ill with notions such as intention and free will (Searle, 1992).

Moreover, the fact that 'reflexes' are not normally introspectively available can explain why human conversation as a research area is much deeper than it seems, and much more challenging than expected.

6.8. How to analyze conversation phenomena

Some studies of back-channels have been motived initially, not by a deep interest in the phenomenon itself, but in the belief that study of back-channels would illustrate the value of a theoretical stance or research method of more general use (Yngve, 1970; Duncan and Fiske, 1985; Schegloff, 1982; Ward, 1997b). As a result, various schools of thought have invested a fair amount of intellectual capital in preparations for the study of such phenomena. It is therefore not surprising that many discussions of methodology for the study of conversation phenomena are heavy in attacks on the assumptions and techniques of workers in different schools (Schegloff, 1982; O'Connell et al., 1990; Searle, 1992; Zimmerman, 1993).

Lacking well thought-out opinions of our own, this section merely recounts which techniques worked for us over the course of our study. We hope that this may contribute indirectly to the eventual resolution of the larger issues.

First, we decided to start from raw data, rather than starting with a theoretical framework. This was because we wanted to study perception and action together, without hypothesizing any intervening variables or representations or abstract structural descriptions. We chose to take the perspective of one participant and consider what his task is, rather than studying the resulting conversation or its structure or rules per se, primarily since we wanted to build a system to take one side of a conversation. Incidentally, our motivation for this was the idea that more natural turn-taking, including back-channels, may be of value in spoken language systems, since it is not always true that people talking to machines desire the interaction to be mechanical and rigid (Johnstone et al., 1995).

Then we began recording conversations. Many researchers interested in prosody choose to work with clean, manageable data, such as prepared sentences read by a radio announcer, monolog 'discourses', and dialogs limited to the exchange of requests and factual information. One might say that this represents a strategy of excluding performance phenomena, in particular, those involving the existence of an interlocutor, those involving the need to decide what to say, and those involving the need to respond in time. Doing so has many advantages, including making results easier to replicate. Working with messy, uncontrolled data might have been impossibly complex and confusing, but we have not found it so in practice. And, with a large enough corpus, the problem of non-replicatable results becomes less serious.

Not knowing where the cues we sought would occur relative to the back-channels, we chose to work with continuous speech. That is, we did not adopt the common practice of pre-segmenting the input into 'fragments' or 'turns', nor the presupposition that a prosodic cue would necessarily be found in the last syllable of the last fragment before the back-channel (Koiso et al., 1998).

We decided to use statistical analysis, reasoning that, whatever the prosodic cue might be, it was probably not something evident to the unaided ear, or it would already have been found. (This hunch turned out to be correct; the low pitch region turns out to be less salient than most of the other factors discussed in section 5) We accordingly gathered statistics on energy, pitch, pitch slope, and so on, and blindly computed correlations between these and the subsequent appearance of back-channel feedback. No correlation seemed to be strong enough to be the cue we sought, so we abandoned this effort. Much later, after discovering the low pitch cue by other means, we took another look at our print-out of correlations, and found that the best correlation was a low pitch point 400 milliseconds before the onset of back-channel feedback in Japanese – had we pursued this at the time, we might have saved ourselves a lot of effort.

We then started to study the conversations in detail. To do this, we did not want to rely on transcripts, even if enhanced with labels indicating perceived intonational features, because we wanted to be able to listen to the sound itself. We did not want to work with tape-recorded or video-recorded data, since we wanted access to the

speech signal at a fine grain, so that we could select any desired small portion of the discourse and listen to it repeatedly. We therefore built a computer program, 'didi', to let us do this, and to let us add transcription labels and notes freely. This program also displays the waveform and its pitch, letting us use both eye and ear; we found this helpful.

Trying to find the cue, we wrote many little bits of software to generate various statistical analyses and graphs. Most of this led nowhere, but one of these graphs, a superimposition of pitch environments in the context of several cases of back-channel feedback, was what led to the discovery of the role of low pitch regions.

This initial analysis was focused on one Japanese conversation, chosen since back-channels were frequent there and since it was clearly an interesting conversation to both parties. We then examined the rest of our corpus to gauge the generality of the low pitch rule.

We were from the start obsessed with objective rules. This led us to strive for careful definition of back-channel feedback, a signal-based (not perceptual) definition of the low pitch cue, and an objective criterion for successful back-channel prediction. These things, together with a largish corpus, allowed us to evaluate different versions of the rule in search of the best one, to compare the strength of this cue to that of other cues, to quantify some differences between Japanese and English, and to identify and direct listening effort to cases where the low pitch cue fails.

Later, in perhaps an unusual research manouevre, after having fixed on a 110 millisecond region of low pitch as a significant prosodic feature by statistical analysis of conversation behavior, we turned to the question of its meaning. By listening to the contexts of the low pitch regions and by computing statistics on what lexical items fall most frequently within low pitch regions, we were able to characterize the communicative functions where low pitch regions occur.

7. Summary

We have found that 110 millisecond regions of low pitch by themselves are a fairly good predictor of subsequent back-channel feedback, and that more obvious factors, such as utterance end, rising intonation, and specific lexical items, account for less than they intuitively seem to. We have also found enough variety in context of occurrence to suggest that there can be no parsimonious meaning-based account of when back-channel feedback appears.

We have also found that speaker-produced cues can account for about half of the occurrences of back-channels, suggesting that back-channel feedback is not simply produced whenever the listener feels like it, but is often cued, encouraged, or allowed by the speaker.

These conclusions apply to both English and Japanese.

Clearly, we have raised more issues than we have resolved.

References

Araki, M. and Ichikawa, A., 1997. Danwa tagu wakingu gurupu katsudo hokoku (Progress report of the discourse tagging working group). 18th spoken language and discourse workshop notes (SIG-SLUD-18), Japan Society for Artificial Intelligence: 31–36.

Boyle, E.A., A.H. Anderson and A. Newlands, 1994. The effects of visibility on dialogue and performance in a cooperative problem solving task. Language and Speech 37: 1–20.

Clancy, P.M., S.A. Thompson, R. Suzuki and H. Tao, 1996. The conversational use of reactive tokens in English, Japanese and Mandarin. Journal of Pragmatics 26: 355–387.

Clark, H.H., 1996. Using language. Cambridge: Cambridge University Press.

Cook, H.M., 1992. Meanings of non-referential indexes: A case study of the Japanese sentence-final particle *ne*. Text 12: 507–539.

Dhorne, F., 1980. Furansugo ni okeru aizuchi no kino (The function of back-channel feedback in French). Nihongogaku 7(12): 38–45.

Dittmann, A.T. and L.G. Llewellyn, 1967. The phonemic clause as a unit of speech decoding. Journal of Personality and Social Psychology 6: 341–349.

Drummond, K. and R. Hopper, 1993. Back channels revisited: Acknowledgment tokens and speakership incipiency. Research on Language and Social Interaction, 26: 157–177.

Duncan, Jr., S. and D.W. Fiske, 1985. The turn system. In: S. Duncan, Jr. and D.W. Fiske, eds., Interaction structure and strategy, 43–64. Cambridge: Cambridge University Press.

Erickson, F., 1979. Talking down: Some cultural sources of miscommunication in interracial interviews. In: A. Wolfgang, ed., Nonverbal behavior: Applications and cultural implications, 99–126. New York: Academic Press.

Ford, C.E., B.A. Fox and S.A. Thompson, 1996. Practices in the construction of turns: The 'TCU' revisited. Pragmatics 6: 427–454.

Ford, C.E. and S.A. Thompson, 1996. Interaction units in conversation: Syntactic, intonational, and pragmatic resources for the management of turns. In: E. Ochs, E.A. Schegloff and S.A. Thompson, eds., Interaction and grammar, 134–184. Cambridge: Cambridge University Press.

Fox, B.A., M. Hayashi and R. Jasperson, 1996. Resources and repair: A cross-linguistic study of syntax and repair. In: E. Ochs, E.A. Schegloff and S.A. Thompson, eds., Interaction and grammar, 185–237. Cambridge: Cambridge University Press.

Fox Tree, J.E. and H.H. Clark, 1997. Pronouncing 'the' as 'thee' to signal problems in speaking. Cognition 62: 151–167.

Fukuchi, Y., 1997. Yokuyo joho ni yoru terefon shoppingu no aizuchi to hatsuwaken kotai no kenkyuu (A study of the relation between prosody and back-channel feedback and turn-taking in shopping over the telephone). B.S. Thesis, University of Tokyo.

Furo, H., 1997. Turn-taking in Japanese conversation: Grammar, intonation, and pragmatics. In: Akatsuka, N., H. Hoji, S. Iwasaki, S.-O. Sohn and S. Strauss, eds., Japanese/Korean linguistics, volume 7, 41–57. Stanford: CSLI Publications.

Gardner, R., 1997. The conversation object *mm*: A weak and variable acknowledging token. Research in Language and Social Interaction 30: 131–156.

Goodwin, C., 1986. Between and within: Alternative sequential treatments of continuers and assessments. Human Studies 9: 205–217.

Hayashi, M., 1996. Toward a proper characterization of the minimal vocalization *u::n* and the like in Japanese conversation: A sketch of some orderliness. Manuscript.

Hayashi, R., 1996. Cognition, empathy and interaction: Floor management of English and Japanese conversation. Norwood, NJ: Ablex.

Hirose, K., H. Fujisaki and S. Seto, 1992. A scheme for pitch extraction of speech using autocorrelation function with frame length proportional to the time lag. 1992 IEEE International Conference on acoustics, speech and signal processing I: 149–152.

Hirschberg, J. and G. Ward, 1995. The interpretation of the high-rise question contour in English. Journal of Pragmatics 24: 407–412.

Horiguchi, S., 1997. Nihongo kyoiku to kaiwa bunseki (Japanese conversation by learners and native speakers). Tokyo: Kuroshio.

Jefferson, G., 1984. Notes on a systematic deployment of the acknowledgement tokens 'yeah' and 'mm hm'. Papers in Linguistics 17: 197–216.

Johnstone, A., U. Berry, T. Nguyen and A. Asper, 1995. There was a long pause: Influencing turn-taking behaviour in human–human and human–computer dialogs. International Journal of Human–Computer Studies 42: 383–411.

Jurafsky, D., R. Bates, N. Coccaro, R. Martin, M. Meteer, K. Ries, E. Shriberg, A. Stolcke, P. Taylor, C. Van Ess-Dykema, 1997. Automatic detection of discourse structure for speech recognition and understanding. Proceedings of the 1997 IEEE Workshop on Automatic Speech Recognition and Understanding.

Jurafsky, D., E. Shriberg, B. Fox and T. Curl, 1998. Lexical, prosodic, and syntactic cues for dialog acts. Association for Computational Linguistics, Workshop on Discourse Relations and Discourse Markers: 114–120.

Kawamori, M., A. Shimazu and K. Kogure, 1994. Roles of interjectory utterances in spoken discourse. 1994 International Conference on Spoken Language Processing: 955–958.

Koiso, H., Y. Horiuchi, S. Tutiya and A. Ichikawa, 1995. Kai hatsuwa tani no onseiteki tokucho to 'aizuchi' to no kanren ni tsuite (The acoustic properties of 'subutterance units' and their relevance to the corresponding follow-up interjections in Japanese). AI Symposium '95 (SIG-J-9501-2), Japanese Society for Artificial Intelligence: 9–16.

Koiso, H., Y. Horiuchi, S. Tutiya, A. Ichikawa and Y. Den, 1998. An analysis of turn-taking and backchannels based on prosodic and syntactic features in Japanese map task dialogs. Language and Speech 41: 295–321.

Ladd, D.R., 1996. Intonational phonology. Cambridge: Cambridge University Press.

LoCastro, V., 1987. Aizuchi: A Japanese conversational routine. In: L.E. Smith, ed., Discourse across cultures, 101–113. New York: Prentice-Hal.

Maynard, S.K., 1989. Japanese conversation. Norwood, NJ: Ablex.

Maynard, S.K., 1997. Analyzing interactional management in native/non-native English conversation: A case of listener response. International Review of Applied Linguistics in Language Teaching 35: 37–60.

Mizuno, Y., 1988. Chugokugo no aizuchi (Back-channel feedback in Chinese). Gengo 7(12): 18–23.

Mizutani, N., 1984. Nihongo kyooiku to hanashikotoba no jittai: Aizuchi no bunseki (Analysis of Back-channels). In: Kindai-ichi haruo hakase koki kinen ronbunshuu: Dai-ni-maku, 261–279. Tokyo: Sanseido.

Mizutani, N., 1988. Aizuchiron (On aizuchi). Nihongo-gaku 7(12): 4–11.

Mizutani, N., 1993. Kyowa kara danwa e (From co-construction to conversation). Nihongo gaku 12(4): 4–10.

Mizutani, O., 1981. Japanese: The spoken language in Japanese life. Tokyo: The Japan Times.

Mueller, F.E., 1996. Affiliating and Disaffiliating with Continuers: Prosodic. Aspects of Recipiency. In: E. Cooper-Kuhlen and M. Setting, eds., Prosody in conversation: Interactional Studies. 113–176. Cambridge: Cambridge University Press.

Noguchi, H., H. Koiso, Y. Fukuda and Y. Den, 1998. Riritsu joho ni motozuita aizuchi sonyu kasho no suitei (Infering back-channel locations from prosodic information). Proceedings of the 4th Annual Meeting of the (Japanese) Association for Natural Language Processing: 484–487.

Novick, D.G. and S. Sutton, 1994. An empirical model of acknowledgement for spoken-language systems. Proceedings 32nd Association for Computational Linguistics: 96–101.

O'Connell, D.C., S. Kowal and E. Kaltenbacher, 1990. Turn-taking: A critical analysis of the research tradition. Journal of Psycholinguistic Research 19: 345–373.

Okada, M., 1996. How the length and pitch of aizuti 'back-channel utterances' and the nature of the speech activity determine preference structure in Japanese. Proceedings of the Twenty-Second Annual Meeting of the Berkeley Linguistics Society: 279–289.

Okato, Y., K. Kato, M. Yamamoto and S. Itahashi, 1996. Riritsu pataan no ninshiki o mochiita aizuchi sonyu to sono hyouka (Prosodic pattern recognition of insertion of interjectory responses and its evaluation). 10th Spoken Language Information Processing Workshop Notes (SIG-SLP-10), Information Processing Society of Japan: 33–38.

Oreström, B., 1983. Turn-taking in English conversation. Lund: Liber Förllag. (Also distributed by Chartwell-Bratt, Bromley, Kent).

Owens, F.J., 1993. Signal processing of speech. New York: McGraw.

Rosenfeld, H.M., 1978. Conversational control functions of nonverbal behavior. In: A.W. Siegman and S. Feldstein, eds. Nonverbal behavior and communication, 291–328. Hillsdale, NJ: Erlbaum.

Rosenfeld, H.M. and M. Hancks, 1980. The nonverbal context of verbal listener responses. In: M.R. Key ed., The relationship of verbal and nonverbal communication, 193–206. The Hague: Mouton.

Sacks, H., E.A. Schegloff and G. Jefferson, 1974. A simplest systematics for the organization of turn-taking for conversation. Language 50: 696–735.

Saito, M., 1997. 'Giji gimon intoneshon' ni kansuru kosatsu (A study of 'mock-question intonation'). B.S. thesis, Tokyo University of Foreign Studies.

Schegloff, E.A., 1982. Discourse as an interactional achievement: Some uses of 'uh huh' and other things that come between sentences. In: D. Tannen ed., Analyzing discourse: Text and talk, 71–93. Washington, DC: Georgetown University Press.

Schiffrin, D. 1987. Discourse markers. Cambridge: Cambridge University Press.

Schuetze-Coburn, S., M. Shapley and E.G. Weber, 1994. Units of intonation in discourse: A comparison of acoustic and auditory analyses. Language and Speech 34: 207–234.

Searle, J.R., 1992. Conversation. In: H. Parret and J. Verschueren, eds., (On) Searle on conversation, 8–29. Amsterdam: Benjamins.

Shriberg, E., R. Bates and A. Stolcke, 1997. A prosody-only decision-tree model for disfluency detection. Eurospeech 1997: 2383–2386.

Siegman, A.W., 1976. Do noncontingent interviewer mm-hmms facilitate interviewee productivity? Journal of Consulting and Clinical Psychology 44: 171–182.

Strauss, S. and Y. Kawanishi, 1996. Assessment strategies in Japanese, Korean, and American English. In: Akatsuka, N., S. Iwasaki and S. Strauss, eds., Japanese/Korean linguistics, vol. 5: 149–165. Stanford: CSLI Publications.

Sugito, M., 1994. Nihonjin no koe (The speech of the Japanese people). Tokyo: Izumi Shoin.

Sugito, M., 1997. Shizen na taiwa ni okeru hibunponteki na hatsuwa no purosodi to kikite no rikai (The prosody of ungrammatical utterances and the listener's understanding in natural conversation). In: Sugito, M., ed., Bunpo to onsei (Speech and grammar), 281–297. Tokyo: Kuroshio.

Takubo, Y., 1994. Towards a performance model of language. First Spoken Language Information Processing Workshop Notes (SIG-SLP-1), Information Processing Society of Japan, 15–22.

Tanaka, H., in press. The particle *ne* as a turn-management device in Japanese conversation. Journal of Pragmatics, to appear.

Tannen, D., 1990. You just don't understand: Men and women in conversation. New York: Morrow.

Ward, N., 1996. Using prosodic clues to decide when to produce back-channel utterances. International conference on spoken language processing: 1728–1731.

Ward, N., 1997a. Inferring processing times from a corpus of conversations. Handout for the First Conference on Computational Psycholinguistics.

Ward, N., 1997b. Responsiveness in dialog and priorities for language research. Systems and Cybernetics 28: 521–533.

Ward, N., 1998. The relationship between sound and meaning in Japanese back-channel grunts. Proceedings of the 4th Annual Meeting of the (Japanese) Association for Natural Language Processing: 464–467.

Ward, N., 1999. Low-pitch regions as dialog signals? Evidence from dialog-act and lexical correlates in natural conversation. In: M. Swerts and J. Terken, eds., ESCA Workshop on Dialog and Prosody, 83–88. Eindhoven: TUE-IPO.

Ward, N. and W. Tsukahara, 1999. A responsive dialog system. In: Y. Wilks, Machine conversations, 169–174. Berlin: Springer.

White, S., 1989. Backchannels across cultures: A study of Americans and Japanese. Language in Society 18: 59–76.

Yamada, H., 1992. American and Japanese business discourse: A comparison of interactional styles. Norwood, NJ: Ablex.

Yngve, V., 1970. On getting a word in edgewise. Papers from the Sixth Regional Meeting of the Chicago Linguistic Society: 567–577.

Zimmerman, D.H., 1993. Acknowledgment tokens and speakership incipiency revisited. Research in Language and Social Interaction 26: 179–194.

Nigel Ward received a Ph.D. from the University of California at Berkeley in 1991. He is interested in human-computer interaction in spoken dialog systems.

Wataru Tsukahara hopes to receive a Ph.D. from the University of Tokyo in 2000. His thesis research focuses on the construction of a 'responsive' and 'sensitive' spoken language system — responsive in the sense of taking turns with the same swiftness as people, and sensitive in the sense of adjusting responses to subtle shifts in the user's attitude as revealed by his utterance timing and prosody.