PROSODIC FEATURES THAT LEAD TO BACK-CHANNEL FEEDBACK IN NORTHERN MEXICAN SPANISH[*]

ANAÍS G. RIVERA and NIGEL G. WARD
University of Texas at El Paso

In order to demonstrate attentiveness during a conversation it is generally necessary for the listener to provide back-channel feedback. To some extent, the times when back-channel feedback is welcome are determined by the speaker and conveyed to the listener with prosodic cues. In this study we sought to identify the cues used for this purpose in Northern Mexican Spanish. Based on quantitative analysis of a corpus of unstructured conversations, we found three cues, of which the most common is a pitch downslope followed by a pitch rise accompanied by a rate reduction on the last syllable and a drop in energy leading to a slight pause.

1. BACK-CHANNELS IN CONVERSATION.

To conduct an engaging conversation it is necessary for the listener to provide feedback. Typically the most common type of feedback consists of a short utterance such as *uhm*, o*k* or *yeah* in English or *ajá*, *si*, *uhm* in Spanish, produced during the turn of the other speaker that encourages the speaker to continue speaking and gives reassurance that the listener is interested. These back-channels (also called known as response tokens, reactive tokens, minimal responses and continuers) are important; lack of back-channel feedback can cause a listener to appear cold, disapproving or rigid.

Such problems are not uncommon in intercultural interactions. For example, it has been reported that the differences in back-channel style between English and Spanish speakers can lead native English speakers to perceive native Spanish speakers as overly aggressive and emotional, and conversely, to lead native Spanish speakers to feel that English native interlocutors are apathetic and cold (Berry 1994). Thus there is practical value to identifying the rules underlying the common patterns of back-channel use.

The positions where the listener produces back-channels depend on both the listener himself and on the speaker. In large part the listener-dependent factors reflect the semantics and pragmatics of the interaction; a listener may choose to demonstrate agreement, understanding, interest, surprise or another emotion in response to the information being conveyed by the speaker. On the other hand, the speaker-dependent factors involve not only the semantic and pragmatic dimensions, but also turn-taking signals, whereby the speaker indicates with prosodic cues what sort of contribution is expected from the listener and when.

2. CORPUS.

The corpus used consisted of five informal conversations between northern Mexican Spanish speakers, all from the state of Chihuahua: five from Chihuahua City, two from Delicias and one from Balleza (Acosta 2004). Two of the dialogs were between two women, two were between two men and one was mixed. The speakers were all in their early twenties. The speakers were

[*] We thank the National Science Foundation for support under grant IIS-0415150, DARPA, Luis Hector Acosta and Jon Amastae.

recorded in situations where they were at ease, and were given no specific instructions. The dialogs seemed fairly natural, with topics including daily life, sports, school, work and fun activities. Each speaker was recorded on a separate channel. The conversations totaled 41 minutes.

The first step in analysis was the identification of the back-channels present in the corpus. This was done fairly casually, but difficult cases were decided according to the definition of Ward and Tsukahara (2000). Thus, to count as a back-channel an utterance had to: 1) respond directly to the content of an utterance of the other, 2) be optional, 3) not require acknowledgement by the other. The initial labeling was done independently by two labelers, both native Spanish speakers. Agreement was reasonable but not high, in part because of cases which were long enough to be ambiguous between back-channels and full turns. To ensure consistency, the corpus was then re-labeled, taking into consideration the opinions of both of the original labelers: this gives the set of actual back-channels. We also labeled possible backchannels; these were places where a back-channel seemed to be invited but did not actually occur. This was done for two reasons. First, since back-channel behavior varies among listeners, we felt that including these places would give a more complete picture, rather than only examining the places where the interlocutor in the corpus actually happened to produce a back-channel. Second, since the corpus consisted of face-to-face dialogs, in some cases back-channel feedback may have been expressed with a head nod or a gesture although a verbal response would also have been appropriate.

TABLE 1:
THE MOST COMMON BACK-CHANNELS IN THE CORPUS, REPRESENTED USING STANDARD ORTHOGRAPHIC CONVENTIONS.

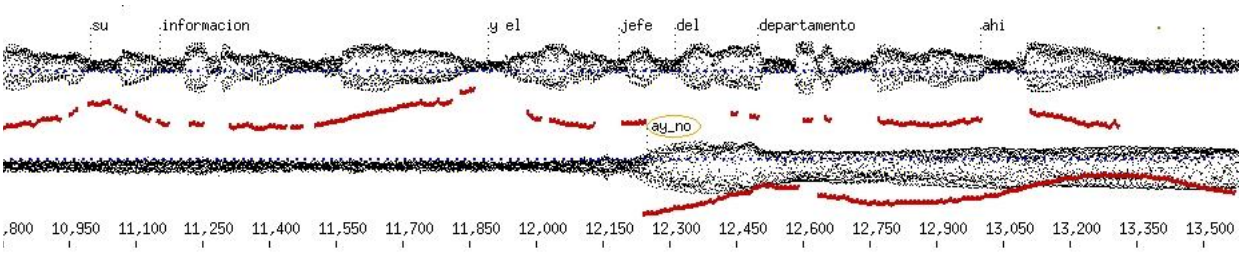
Rank	Back-channel	Number
1	si	74
2	si si	24
3	ajá	21
4	mjm	13
5	laughter	12
6	ei	11
7	no	9
8	mm	7
9	ah	3
10	ay no	2

There were 195 actual occurrences of backchannel feedback; thus a back-channel occurred on average every 13 seconds. There were also 124 possible back-channel points. A variety of sounds served as back-channels; the most common are shown in Table 1. The other backchannels seen were mostly multi-word combinations of these. Phonetic labeling was not done, however it is worth noting that the vowels in *ajá* are close to a schwa, and the letter <j> represents a back fricative. Semantic labeling was also not done, however there was clearly substantial variation in the nuances being conveyed; various tokens conveying greater or lesser degrees of energy, interest, amusement, agreement, sympathy, surprise, and approval. Figure 1 illustrates a back-channel from the corpus. This dialog fragment came after the speaker said:

Ya, o sea, estaba en su maquina restaurando su información y el jefe del departamento ahí…
'[He] was at his machine recovering his data and the department head was there...'

and the listener is expressing sympathy. This example is somewhat unusual in that the backchannel is longer than a second and almost completely overlaps the speaker's continued turn, however in other respects it is typical. The audio for this and other examples is available at http://www.cs.utep.edu/isg/members/anais/

FIGURE 1. THE PROSODIC CONTEXT OF A BACK-CHANNEL DEMONSTRATING SYMPATHY.
Here the upper track is the speaker and the bottom track the listener. Each track includes, from top to bottom, the transcription, the signal, and the pitch contour in log scale.



3. ANALYSIS.

The aim of the analysis was to identify prosodic cues from the speaker that cue (or invite) the listener to produce back-channel feedback. For this we used an eclectic method (Ward and Al Bayyari 2006) that has earlier proved successful with other languages. The key strategy was to find one or more prosodic patterns that occur frequently before back-channels, but infrequently in other contexts.

It quickly became apparent that there is no pitch pattern common to all cases, and thus no simple rule for determining when the speaker is cuing the listener to produce a back-channel. In addition many of the prosodic patterns common before back-channels were also frequently present elsewhere in the dialogs.

Analysis proceeded by a process of hypothesis formulation and refinement. After we had an idea of what the prosodic pattern was, we formalized it and then incorporated it in the system as a predictive rule. We could then use it to predict backchannel occurrences, and we could see (and hear) whether and how these predictions did or did not match backchannel responses in the corpus. For each hypothesized rule, we examined correct predictions, missing predictions, and false predictions, and then used this to refine the rule, typically by incorporating additional features. This iterative process led to the discovery of three patterns that significantly precede back-channel behavior.

Finally, to obtain the best possible quantitative description of each pattern, we systematically varied the parameters to find the description that gave the best performance. Here the job for the rule was, given the prosodic information in one track of a dialog, to predict where in the other track the back-channels occurred. The metric of performance was the F-measure, that is, the harmonic mean, of the accuracy (the percent of the predictions that matched a

backchannel) and the coverage (the percent of backchannels that matched a prediction) (Ward and Tsukahara 2000).

4. RESULTS.

So far we have identified three common prosodic patterns preceding back-channels, that is, three prosodic cues. The first and most common consists of a low pitch region followed by a rise in pitch accompanied by a reduction in rate. The second consists of a flat, low pitch region. The third group consists of a steep pitch drop, usually an indicator of an amusing comment. The rest of this section discusses each in turn.

4.1 Low-High-Slow Pattern. The most common cue is characterized by a pitch downslope or low region followed by a pitch rise accompanied by a rate reduction on the last syllable and a drop in energy leading into a slight pause. Best performance is obtained with a rule modeling the listener as producing a backchannel 200ms after an utterance by the speaker including:

- a low pitch region lasting for at least 50ms and for no more than 200ms with the pitch continuously below the 28th pitch percentile for that speaker, followed by
- a pitch rise ending above the 75th pitch percentile for that speaker, and lasting at least 80ms and no more than 300ms, and including or followed within 200 ms by
- a lengthened vowel (one lasting at least 100ms), followed within 80 ms by
- a period of silence lasting at least 200 ms.

Figure 2 shows this in diagram form. Figure 3 is an example where the speaker produces this intonation pattern and the listener responds with a back-channel. In this example, the speakers are discussing vacation plans and the back channel occurs after the speaker says:

> *El martes que la ví...*
> 'On Tuesday when I saw her… [making reference to a common friend]'

This rule gives 28.7% coverage, thus it explains over a quarter of the occurrences of back-channels and back-channel opportunities in the corpus. The accuracy is 14.2%, meaning that it over-predicts significantly, although this is far better than the baseline, namely the 6.1% accuracy expected by random guessing.

FIGURE 2: DIAGRAM OF THE LOW-HIGH-SLOW PATTERN.

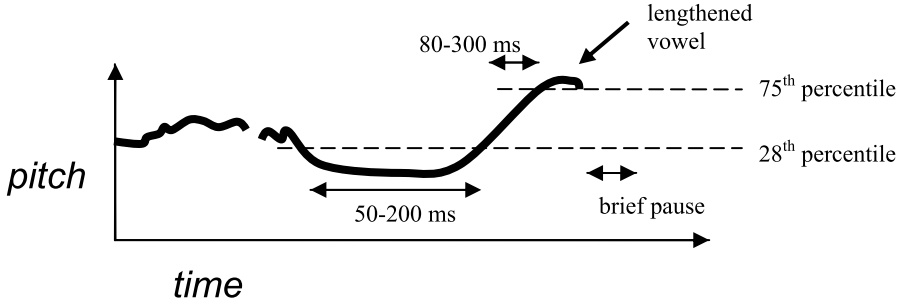
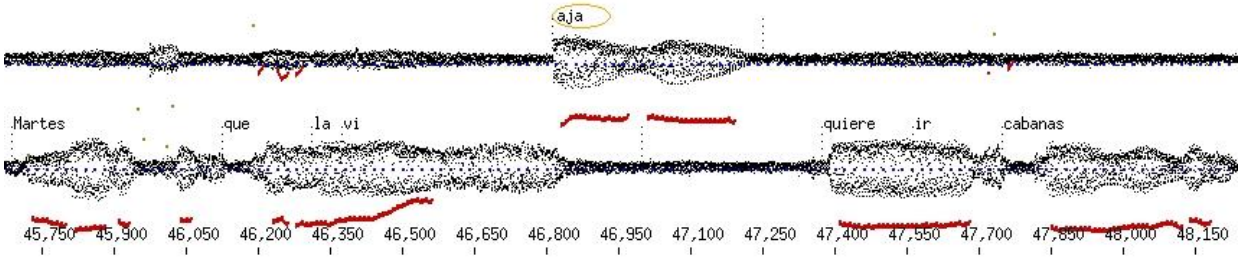


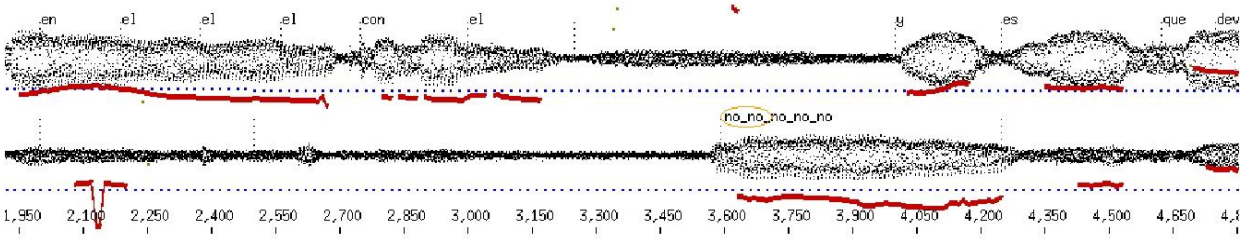
FIGURE 3. DIALOG FRAGMENT ILLUSTRATING THE LOW-HIGH-SLOW BACK-CHANNEL CUE.



4.2 Flat Pitch Region Pattern. The second cue for back-channels cue is a region of flat pitch. These cues accounted for 7.9% (i.e., the coverage was 7.9%) of both the possible and spoken back-channels but also occurred in many other places giving an accuracy of 6.5%. Figure 4 is an example of this type of cue. In this example the back-channel occurs after the speaker says:

Dos ceros y dos cincos me saque en el el el con el…
'[My grades were] two zeroes and two fives with him…'

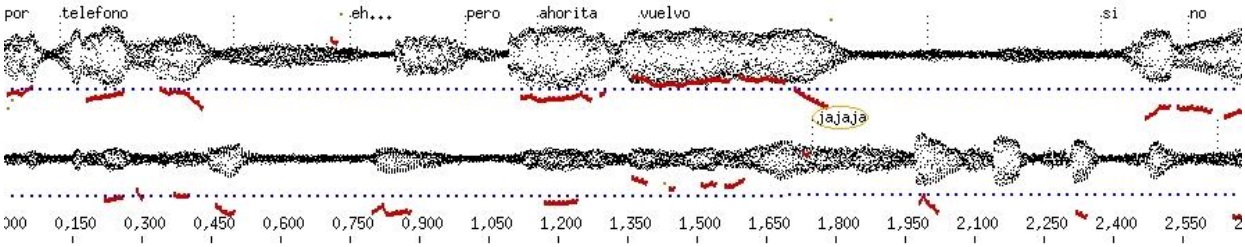
FIGURE 4. A FLAT PITCH REGION AS A BACK-CHANNEL CUE.



4.3 Pitch Drop Pattern. In some cases back-channels are preceded by a pitch drop, especially if the back-channel consists of laughter; indeed, this pitch drop typically marks the punch-line. Pitch-drop based backchannels were more common in the male-male dialogs, and in one dialog where the speakers appeared less focused on the conversation and the back-channels appear in less consistent places. Overall this type of cue gives 71.6% coverage and 7.7% accuracy. Figure 5 gives an example of this cue in a punch-line. In this example the backchannel comes after the speaker says:

Eh voy a hablar por teléfono eh… pero ahorita vuelvo.
'I'm going to make a phone call uhm... but I'll be right back.'

FIGURE 5. A PITCH DROP AS A BACK-CHANNEL CUE.



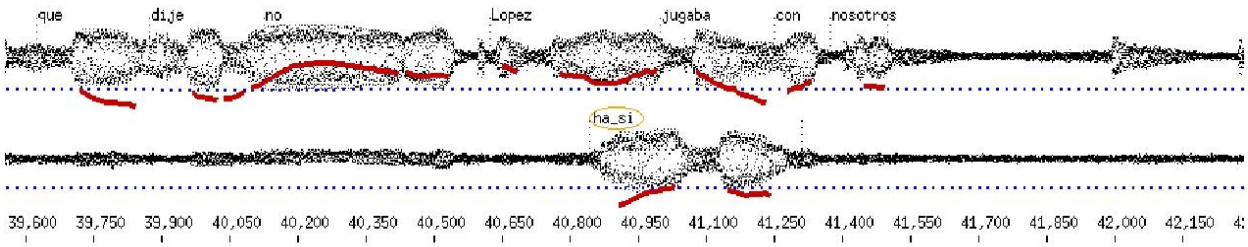
5. ERROR ANALYSIS.

There are cases in which the rules mentioned above failed to predict a back-channel (generated a miss) or predicted a back-channel in places were there was none (generated a false prediction). Those cases represent aspects of back-channeling that our simple three-rule model doesn't account for. Most of these aspects are beyond the scope of this study, including individual differences in back-channeling style and the processes of information transmission and processing. There are four other main causes for these errors.

5.1 Ongoing Speech. The most common cause for misses was the appearance of back-channels overlapping the other's ongoing utterance. Since our most common rule is based on the presence of a small pause, if the speaker continues speaking with no significant pause, this cue is missed. This was not uncommon; as Berry (1994) observes, overlapped speech is common in Spanish. Figure 6 shows an example of a miss due to ongoing speech; here the back-channel occurs in the middle of the sentence:

¿Te dije no? López jugaba con nosotros.
`I told you right? Lopez played [soccer] with us.'

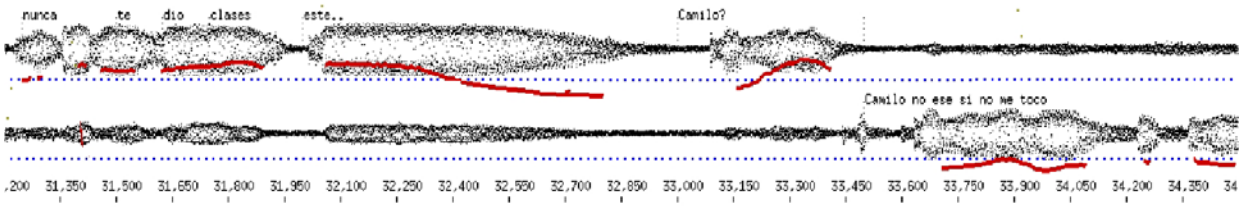
FIGURE 6. OVERLAPPED SPEECH CAUSING A MISS.



5.2 Overlapped Speech. Another cause was cases where both speakers were talking at once. In some such cases there was a back-channel cue produced by one speaker, but of course no back-channel by the other, so these counted as false predictions.

5.3 Questions. Another major cause of false predictions was yes/no questions, whose intonation is similar to that of back-channel cues. Questions are similarly characterized by a rising pitch intonation at the end of a sentence, the main differences being: 1) questions are not as frequently preceded by a lowered pitch region; 2) in questions a final lengthening is less common; and 3) in questions the pitch rise may be very long, sometimes lasting throughout the duration of the utterance. However, there are many exceptions to these tendencies. Figure 7 shows an example of a false prediction due to a case where the prosody of a yes-no question happened to meet the criteria for our first cue.

FIGURE 7. YES-NO QUESTION INTONATION.



5.4 Gender Differences. There are significant differences in performance between conversations, and in general the rule was much less successful on those dialogs with both speakers male (coverage 19%, accuracy 9%) than on the others (coverage 30%, accuracy 18%). This might reflect differences in feedback styles between genders.

6. CONCLUSION.

This paper has shown that in Spanish, as in other languages, the times when back-channels are appropriate are signaled by the speaker to the listener in part by prosodic cues. The specific cues identified have not been seen before; certainly they differ from those seen in English (Ward and Tsukahara 2000). It is intriguing that our prosodic account of when back-channels are appropriate in Spanish has weaker explanatory power, quantitatively, than our account for English, but we do not know yet whether this is a real difference or a mere reflection of the fact that this corpus consists entirely of face-to-face dialogs between friends.

Intercultural dialogs are sometimes awkward, and differences in back-channeling practices seem to be a contributing factor: second language learners may back-channel inappropriately or, perhaps equally undesirably, they may fall back to a more rigid, cold back-channel free listening style. Teaching learners the rules governing back-channeling seems to require many examples and controlled practice of various kinds; we are currently developing a training sequence to do this effectively (Ward et al. 2007).

We have identified these prosodic patterns as cuing back-channels without considering individual differences, without regard to interactions with pitch-accent, micro prosody, or other prosodic functions, and without consideration of how back-channeling and back-channel cuing interacts with various specific dialog activities. Future work should investigate these aspects of the phenomenon.

REFERENCES

ACOSTA, LUIS HECTOR. 2004. Prosodic features that cue back-channel responses in northern Mexican Spanish. Computer Science Department Masters Thesis, University of Texas at El Paso.

BERRY, ANNE. 1994. Spanish and American turn taking styles: A comparative study. In L. F. Boulton, editor, *Pragmatics and Language Learning Monograph Series, Volume 5*, 1994, pages 180-190. University of Illinois, Urbana-Champaign: Division of English as an International Language, 1994.

WARD, NIGEL, RAFAEL ESCALANTE, YAFFA AL BAYYARI, and THAMAR SOLORIO. 2007. Learning to show you're listening. *Computer Assisted Language Learning*, under submission.

WARD, NIGEL and WATARU TSUKAHARA. 2000. Prosodic features which cue back-channel feedback in English and Japanese. *Journal of Pragmatics*, 32:1177-1207.

WARD, NIGEL and YAFFA AL BAYYARI. 2006. A case study in the identification of prosodic cues to turn-taking: Back-channeling in Arabic. In Interspeech 2006 Proceedings.