Cross-cultural Investigation of Prosody in Verbal Feedback in Interactional Rapport

Gina-Anne Levow[1], *Susan Duncan*[2], *Edward T. King*[2]

[1]Department of Computer Science, University of Chicago, Chicago, IL 60611
[2]Department of Psychology, University of Chicago, Chicago, IL 60611

levow@cs.uchicago.edu, deng@uchicago.edu, etking@gmail.com

Abstract

Aspects of speech and non-verbal behavior allow conversational partners to establish and maintain rapport by signaling engagement or endorsement. In the verbal channel, these factors encompass requests for and production of vocal feedback, as well as lexical and grammatical mirroring. However, these cues are often subtle and culture-specific. Here, we present a preliminary investigation of the differences in elicitation and provision of vocal feedback across three diverse language/cultural groups: American English, Gulf/Iraqi Arabic, and Mexican Spanish. We describe our corpus of unrehearsed dyadic story-telling interactions, with listeners who had been instructed to be "active and engaged." Based on a fully-transcribed and aligned subcorpus of 79 interactions, we identify fundamental contrasts in expectations for and production of vocal feedback. We identify dramatic differences in the rates of listener verbal feedback across the groups. However, we find that some significant pitch-related prosodic contrasts are robustly employed across these diverse groups, while we do find differences in the use of other pitch and intensity cues. These differences will inform the development of culturally-sensitive conversational agents, able to engage in more effective dialogue.

Index Terms: rapport, multi-lingual analysis, vocal feedback, prosody

1. Introduction

Members of different cultures may behave quite differently from one another when interacting face-to-face. Culture-specific aspects of speech and non-verbal behavior are signals that enable members of a culture to establish and maintain a sense of rapport with one another over intervals of interaction. Rapport, and the means by which conversation partners achieve it, is important to study systematically, because rapport is known to increase the likelihood of success of goal-directed interactions and also to promote knowledge sharing and learning. Subtle cues signal engagement, endorsement, or appreciation. In the verbal channel, these include mirroring of word choices and of grammatical structures as well as vocal feedback.

The goal of our current research, is to elucidate behavioral similarities and differences among three language/cultural groups – American English, Gulf/Iraqi Arabic, and Mexican Spanish – focusing specifically on how individuals within each culture establish and maintain rapport. The cross-culture comparative dimension of our study potentiates the development of conversational agents that model culturally distinctive behaviors that are related to maintenance of interactional rapport. In this paper, we investigate the elicitation and production of listener verbal feedback as part of the establishment and maintenance of dyadic rapport, focusing on the use of prosodic cues by the speaker that can elicit verbal feedback.

2. Related Work

Fundamental work by [1] provided an analysis of conversational interaction as fundamentally rule-governed, specifying a range of cues that guided turn-taking and feedback. He described a range of multimodal cues, including gaze, posture, nod, and prosody that were cumulative, but also defeasible. The production of verbal and non-verbal feedback by the listener has evoked substantial interest, both within and across cultures. [2, 3] analyzed nodding and other listener feedback behavior in Japanese, while [4, 5] contrasted these behaviors in Japanese, English, and Mandarin Chinese. These studies highlighted differences in the frequency and form of this feedback across languages, with Japanese exhibiting the most frequent feedback followed by Chinese and then English.

Several studies have investigated the verbal, especially prosodic, cues that evoke listener feedback from a more quantitative, computational perspective. Recently, [6] investigated speaker backchannel-inviting cues in human-human dialogue in the Columbia Games Corpus. He found that increased pitch, increased intensity, rising pitch, and voice quality measures as well as certain POS bigram patterns were associated with listener feedback, with additional cues increasing the likelihood of backchannels. [7] employed a shallow processing model using pause duration and POS trigrams to predict backchannels. A series of investigations [8, 9, 10] explored prosodic cues to backchannels across Japanese, English, Arabic, and Spanish. A period of low pitch for English and Japanese was found to be a good predictor, while a region of falling pitch was shown to be a good cue for backchannel in Arabic, and three pitch patterns were suggested as cues to backchannel in Spanish. Our current work aims to produce a larger-scale and more controlled corpus for investigation of cross-language and cross-cultural indicators of social resonance, to which listener verbal feedback contributes.

3. Multi-cultural dyadic rapport corpus

To support the investigation of verbal and non-verbal cues to dyadic rapport, we audio-videotaped dyads from each language/culture, engaged in an unrehearsed story-telling activity. Participants were recruited by advertisement either from the University community or cultural centers serving the particular language/cultural groups. Participants were recruited in pairs from those with existing close friendship or family rela-

	English	Spanish	Arabic
Male-Female	9	10	16
Male-Male	11	5	11
Female-Female	14	14	18
Total	34	29	45

Table 1: Dyad distribution across language/cultural groups

Figure 1: Top-to-bottom: American English-, Iraqi Arabic-, and Mexican Spanish-speaking dyads engaged in the Pear Film elicitation. Listener close-ups are the leftmost stills.

tionships; thus we could ensure existing rapport between conversants. To minimize the influence of other language/cultural experience and background, we excluded subjects who had significant foreign language experience or had lived abroad for an extended period of time.

A participant in the role of Speaker was shown the "Pear Film" [11]. The "Pear Film" is a six minute film with no speech developed at the University of California, Berkeley in 1975 as a general language-independent elicitor. The Speaker then related the story of the "Pear Film" to their partner. We instructed listeners to be "active and engaged" in the story-telling task. They understood that, after hearing the story, they would be videotaped themselves, re-telling it to an investigator.

At this time, our corpus of dyadic discourse data comprises elicitations from 45 Arabic-speaking dyads (Iraqi and Emirati) videotaped in the Chicago area in the United States, in Amman, Jordan and in Al-Ain, the United Arab Emirates, 29 Mexican dyads videotaped in the United States, and 34 American dyads, also videotaped in the United States. A summary of current corpus dyads appears in Table 1.

To support close audio and multi-modal analysis, we employed the following recording configuration. The participants were seated at a table facing each other. Each was fitted with a close-talking, head-mounted microphone, recording each speaker on a single channel. Three video-camera recordings were created: one camera to the side of the table at the midpoint capturing both participants (dyad view) and one facing each participant on the table itself. A clapper was used to support manual synchronization of the three views and creation of a composite trio view from all cameras. The two audio channels were fed directly to the camcorder capturing the dyad view. Example images from dyadic interactions are shown in Figure 1.

	English	Spanish	Arabic
Dur (secs): Mean	359.11	251.89	274.74
Dur (secs): Min	92.45	108.57	112.78
Dur (secs): Max	587.29	559.76	472.77
Dur (Phr): Mean	246.8	114.3	175
Dur (Phr): Min	91	36	45
Dur (Phr): Max	499	350	340
Dur (Wds): Mean	1210	728.1	650.8
Dur (Wds): Min	383	327	192
Dur (Wds): Max	2468	1800	1453

Table 2: Narrative durations in seconds, phrases, and words across language/cultural groups

3.1. Corpus annotation

The corpus is being annotated for both text and multimodal behavior, though we focus here on vocal behavior. Dyad speech was transcribed and manually aligned to silence-delimited phrase boundaries, roughly similar to interpausal units (IPUs) in prior work. In the case of Arabic, all transcripts were also manually vowellized and automatically transliterated into Hans Wehr format, a Latin alphabet-based transliteration system which uses diacritics but no digraphs, for readability and to facilitate downstream processing. This rough alignment was then automatically force-aligned at the word-level using the CUSonic alignment tool [12]. For Spanish and Arabic, we employed the language porting functionality in CUSonic. Annotators reviewed and revised word alignments as necessary. The resulting corrected alignments then constrained a final phone-level forced alignment phase. Annotators also provided a rough translational gloss at the word and phrase levels. To date, this process has yielded 79 fully transcribed and aligned narratives for analysis: 31 English, 25 Spanish, and 23 Arabic.

3.2. Corpus Overview

Overall corpus narrative duration statistics are provided across all language conditions in Table 2. These statistics indicate some substantial variation within and across language/cultural elicitation groups. ANOVA indicates a significant effect ($p < 0.05$) of language/culture on narrative duration. Tukey post-hoc tests show that the English narratives are significantly longer than both the Spanish and Arabic story retellings on both time and phrase duration measures. Differences between Spanish and Arabic dialogue durations do not reach significance. Due to the complexity of Arabic morphology, direct comparisons of word count with other languages are likely to be unreliable.

4. Frequency of Verbal Feedback

Work by [4] highlighted the differences in the frequency of backchannel behavior in Japanese, English, and Mandarin Chinese. Our controlled corpus of narratives collected in a consistent setting with a uniform, language-independent elicitor allows us to compare the rates of listener verbal feedback in American English, Mexican Spanish, and Gulf Arabic. Since the speaker controls the floor throughout the retelling, we treat all listener utterances as instances of vocal feedback in our setting. Given the high variance in narrative duration, we compare the ratio of listener feedback phrases to total utterance phrases in a dyadic narrative across speaker groups.

The contrasts appear in Table 3. By ANOVA, we find a significant effect of language/cultural group on the rate of listener

Feedback Rate	English	Spanish	Arabic
Average	0.2	0.15	0.29
Maximum	0.75	0.41	0.53
Minimum	0.0	0.0	0.019

Table 3: Contrasting rates of listener verbal feedback across language/cultural groups.

verbal feedback ($p < 0.025$). Tukey post-hoc tests indicate significant differences between Arabic and both English and Spanish. However, the differences between English and Spanish do not reach significance. In examining these interactions, the contrasts in degree and frequency of listener verbal feedback are highly salient. In several of the Spanish narratives and some of the English narratives, the listener produces almost no verbal feedback at all, in sharp contrast to the highly interactive Arabic retellings. These impressionistic observations are supported by the quantitative statistics.

5. Prosody and Listener Verbal Feedback

Following prior research which has indicated the utility of pitch and intensity cues, we extract a range of pitch and intensity features to identify those associated with listener vocal feedback in our narrative setting across the three language/cultural groups of interest. For each speaker phrase, we extract the following prosodic measures, computed over the full phrase or utterance and over the final word of the phrase:

- Pitch:
  - Maximum, minimum, mean
  - Slope over last half
- Intensity:
  - Maximum, minimum, mean
  - Slope over last half

All measures are computed using Praat's [13] 'To Pitch...' and 'To Intensity...' functions. Raw values are then log-scaled and z-score normalized by speaker for each measure.

5.1. Contrastive Prosodic Analysis

We find significant prosodic differences in speaker utterances which precede listener verbal feedback and those which do not. We find contrasts at both the word and phrase level, for both pitch and intensity measures. While we find some common broad trends across language/cultural groups, particularly in pitch contrasts, there are some substantial differences.

Specifically, all language/cultural groups exhibit significantly lower normalized measures of pitch height for the final word in the phrase in instances followed by listener vocal feedback than in those without such feedback (t-test, two-tailed, $p < 0.025$). For Arabic speakers, there are decreases in pitch maximum and pitch mean; for American English and Mexican Spanish speakers, there are decreases in pitch mean and pitch minimum. At the phrase level, though, none of the differences in measures of pitch height reach significance for our Arabic speaking subjects, whereas both the English and Spanish speakers exbihit significantly lower normalized pitch mean and pitch minimum, consistent with the word level measures. American English speakers also exhibit a small but significant increase in pitch maximum in phrases preceding listener vocal feedback relative to those with such feedback ($p < 0.005$). These contrasts are shown in Figure 2 for Arabic, Figure 3 for English, and Figure 4 for Spanish.

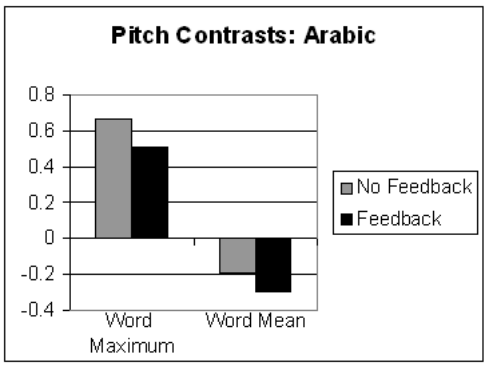

Figure 2: Pitch contrasts cuing verbal feedback: Arabic

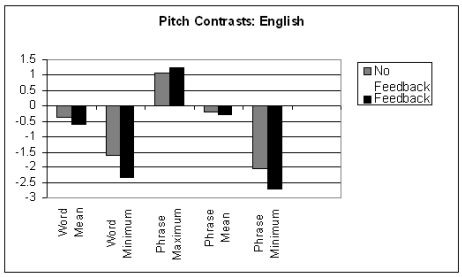

Figure 3: Pitch contrasts cuing verbal feedback: English

In contrast, the use of intensity as a cue for listener vocal feedback is less consistent. Both English and Spanish speakers produce speech of significantly lower intensity ($p < 0.005$) prior to vocal feedback for the measures of intensity mean and maximum on the final word and across the phrase. American English speakers also demonstrate a significant decrease in pitch minimum in the final word of phrases preceding vocal feedback ($p < 0.05$). These contrasts appear in Figure 5 for American English and Mexican Spanish below. However, the only contrast observed in intensity in the Arabic speech is a small but significant ($p < 0.0025$) *increase* in phrasal pitch minimum in contexts before listener feedback. Measures for pitch and intensity slope did not reach significance.

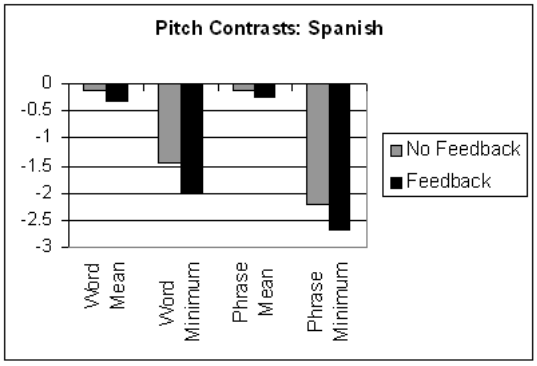

Figure 4: Pitch contrasts cuing verbal feedback: Spanish

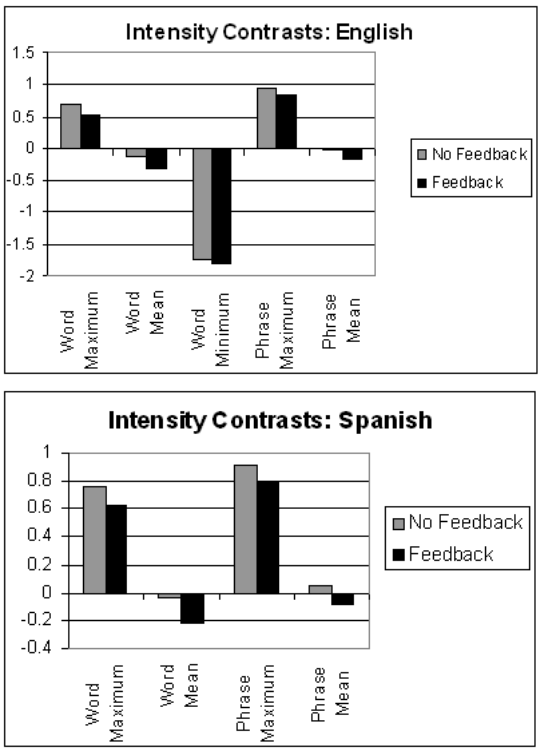

Figure 5: Pitch contrasts cuing verbal feedback: English (top) and Spanish (bottom)

6. Discussion

The contrasts in the rates of verbal feedback suggest that we may be able to extend the ranking of languages based on frequency of listener feedback, as discussed by Maynard. The current findings place Arabic ahead of English and Spanish on a continuum of more to less verbal feedback. However, it remains to be seen whether the vocal feedback is representative of the rate of all forms of listener feedback.

The contrasts in acoustic-prosodic measures associated with vocal feedback are largely consistent with several proposals of low pitch regions as cues to feedback or jump-in points for the listener [8]. However, they stand in significant contrast to the results obtained by [6], where *increased* pitch height and intensity characterized backchannel-inviting utterances. One possible explanation might lie in the differences in interactional setting. In Gravano's work, conversants were engaged in a problem-solving dialogue without visual access; in the cases where low pitch and, in some cases, intensity played a role, the interactions often involved face-to-face narratives, casual conversations, or meetings. These contrasts might contribute to the dramatic differences in prosodic cue usage.

7. Conclusion & Future Work

We are creating a corpus of dyadic discourse data to investigate cross-language/cross-culture differences in multimodal establishment and maintenance of rapport by speakers of American English, Mexican Spanish, and Gulf/Iraqi Arabic. The current aligned and transcribed corpus subset has allowed us to identify significant differences in the rate of listener verbal feedback across these groups and to investigate the role of some prosodic cues in elicitation of this feedback, highlighting a robust use of pitch information.

As additional annotations of multimodal cues become available, it will be possible to assess the interaction of prosodic and non-verbal cues in elicitation of both verbal and non-verbal listener feedback. This information will allow the investigation of whether groups which produce lower rates of verbal feedback compensate with feedback in other modalities. We will also employ both verbal and non-verbal cues to predict and generate appropriate signals to maintain rapport in interaction across cultures.

8. Acknowledgements

This work was supported by NSF BCS 0729515. We also thank the team of transcriber/analysts that made this work possible.

9. References

[1] S. Duncan, "Some signals and rules for taking speaking turns in conversations," *Journal of Personality and Social Psychology*, vol. 23, no. 2, pp. 283–292, 1972.

[2] S. Kita and S. Ide, "Nodding, aizuchi, and final particles in Japanese conversation: how conversation reflects the ideology of communication and social relationships," *Journal of Pragmatics*, vol. 39, pp. 1242–1254, 2007.

[3] M. Kogure, "Nodding and smiling in silence during the loop sequence of backchannels in Japanese conversation," *Journal of Pragmatics*, vol. 39, pp. 1275–1289, 2007.

[4] S. Maynard, "Conversation management in contrast: listener response in Japanese and American English," *Journal of Pragmatics*, vol. 14, pp. 397–412, 1990.

[5] P. Clancy, S. Thompson, R. Suzuki, and H. Tao, "The conversational use of reactive tokens in English, Japanese, and Mandarin," *Journal of Pragmatics*, vol. 26, pp. 355–387, 1996.

[6] A. Gravano and J. Hirschberg, "Backchannel-inviting cues in task-oriented dialogue," in *Proceedings of Interspeech 2009*, 2009, pp. 1019–1022.

[7] N. Cathcart, J. Carletta, and E. Klein, "A shallow model of backchannel continuers in spoken dialogue," in *Proceedings of the tenth conference on European chapter of the Association for Computational Linguistics - Volume 1*, 2003, pp. 51–58.

[8] N. Ward and W. Tsukuhara, "Prosodic features which cue back-channel responses in English and Japanese," *Journal of Pragmatics*, vol. 32, no. 8, pp. 1177–1207, 2000.

[9] N. Ward and Y. Al Bayyari, "A prosodic feature that invites back-channels in Egyptian Arabic," *Perspectives in Arabic Linguistics*, 2007.

[10] A. Rivera and N. Ward, "Three prosodic features that cue back-channel in Northern Mexican Spanish," University of Texas, El Paso, Tech. Rep. UTEP-CS-07-12, 2007.

[11] W. Chafe, "The Pear Film," 1975. [Online]. Available: http://www.linguistics.ucsb.edu/faculty/chafe/pearfilm.htm

[12] B. Pellom, W. Ward, J. Hansen, K. Hacioglu, J. Zhang, X. Yu, and S. Pradhan, "University of Colorado dialog systems for travel and navigation," 2001.

[13] P. Boersma, "Praat, a system for doing phonetics by computer," *Glot International*, vol. 5, no. 9–10, pp. 341–345, 2001.