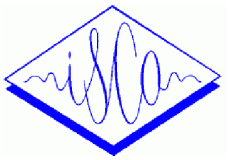
ISCA Archive
http://www.isca-speech.org/archive



INTERSPEECH 2012
ISCA's 13th Annual Conference
Portland, OR, USA
September 9-13, 2012

Contrasting Cues to Verbal and Non-Verbal Backchannels in Multi-lingual Dyadic Rapport

Gina-Anne Levow[1], *Susan Duncan*[2]

[1]Department of Linguistics, University of Washington, Seattle, WA, USA
[2]Department of Psychology, University of Chicago, Chicago, IL, USA

levow@u.washington.edu, deng@uchicago.edu

Abstract

Diverse multi-modal behaviors provide important cues in establishing and maintaining interactional rapport. However, these behaviors are often subtle and culture-specific. In this paper, we focus on two forms of backchannel behavior: vocal backchannels and non-verbal headnods. We employ a corpus of quasi-monologic story-telling interactions elicited from three distinct language/cultural groups: American English, Mexican Spanish, and Iraqi Arabic speakers. Through this corpus, we investigate prosodic cues associated with these two different types of verbal feedback. We identify both similarities and differences in the cues exploited by the speakers of these diverse language/cultural groups. Although both typically classed as backchannels, we observe substantial differences in cues associated with verbal backchannels and head nods across these languages. These contrasts argue for a more fine-grained analysis of the use and role of diverse social resonance behaviors.

Index Terms: prosody, backchannels, multilingual analysis, multimodal

1. Introduction

The character of face-to-face interaction can differ significantly from one cultural group to another. For members of a specific cultural group, certain speech and nonverbal behaviors may enable them to establish a sense of rapport with others. Rapport has been shown to increase the success of goal-directed interactions, and it can also promote knowledge sharing and learning. Thus, studying rapport systematically is important. Previous work has identified cross-cultural differences in a variety of behaviors that may play a role in signaling mutual engagement, endorsement or appreciation. These behaviors include nodding [1], posture [2], facial expression [3], gaze [4], cues to vocal back-channel [5, 6, 7], and co-verbal gesturing [8] as well.

[9] argued that coordination, positive emotion, and mutual attention are key elements of interactional rapport. In the verbal channel, coordination is manifested in regulation of turn-taking and backchannels among conversational participants. Foundational work by [10] established that conversational interaction is fundamentally rule-governed. Multi-modal cues including gaze, posture, nod and prosody were shown to signal turn-taking. In addition, [1, 11] contrasted nodding and other listener behaviors in Japanese, English and Mandarin Chinese. These studies highlighted the cross-cultural differences in the type and frequency of listener response behavior in different languages. The Japanese speakers exhibited the most frequent feedback, followed by Chinese and then English.

Several recent studies investigated the role of verbal, especially prosodic, cues in signaling listener verbal feedback in dyadic and multi-party scenarios based on a quantitative and computational perspective. In [12], features from shallow processing, like pause duration and part-of-speech (POS) tag sequences, are shown to be helpful in predicting backchannels. In [13],[14], it was reported that increases in pitch and intensity, as well as certain POS patterns, are key verbal backchannel-inviting cues in task-oriented dialog. The multi-lingual comparison discussed in [5, 6, 7] found that pitch patterns, e.g., periods of low pitch or drops in pitch, are positively associated with listener backchannels in Japanese, English, Arabic and Spanish speakers. [15] presented initial analyses of another multi-modal corpus of American English, Mexican Spanish, and Iraqi Arabic, highlighting significantly greater rates of listener verbal contributions in Arabic-speaking dyads than in either American English or Mexican Spanish dyads. In addition, initial prosodic analysis of contexts eliciting verbal contributions indicated widespread use of pitch and intensity cues. Using the same corpus, [16] demonstrated improved prediction of listener verbal feedback through a combination of class reweighting and oversampling using Support Vector Machines trained on prosodic features.

Other research has investigated cues to and recognition of backchannels in other modalities such as head nod. [17, 18] employed a multi-modal corpus of English dyadic interactions to investigate verbal and non-verbal backchannels, associating differences in backchannel function with differences in backchannel form. Recently, [19] investigated cultural differences in gaze, proxemics, and backchannel behavior in a multi-modal corpus of American English, Mexican Spanish, and Arabic speakers. [20] developed a highly effective technique for predicting backchannel nods, exploiting pause and shallow lexical cues, as well as some prosodic cue patterns in a Conditional Random Field framework. [21] also investigated prediction of head nods using Support Vector Machines based on speaker silence, pitch, and head movement.

While more theoretical work has discussed a broad range of interacting multi-modal cues to interactional rapport, the majority of the empirical studies, such as those above, have focused on a single cue to social resonance, such as verbal backchannels or nods. Furthermore, these sorts of backchannels are often treated as a single monolithic class, although the literature has proposed a more fine-grained characterization of backchannel discourse function, as explored in [18, 22]. Limitations on availability of suitable annotated corpora as well as intrinsic complexity have led to this focus. We exploit a multi-modal, multilingual data corpus of quasi-monologic story-telling among conversational dyads with high interactional rapport. We investigate eliciting cues associated with both listener verbal backchannels and nods across three diverse language/cultural

groups: Iraqi Arabic, American English, and Mexican Spanish speakers. We first analyze the prosodic speaker cues associated with each of these types of listener backchannel independently, identifying similarities and contrasts across languages and backchannel types. In particular, we identify a divergence in the use of prosodic cues to these two types of interactional signals. An integrated analysis allows us to explain this divergence and to provide a more nuanced understanding of the use of these social resonance signals in dyadic conversation.

2. Multi-modal Rapport Corpus

We employ the same multi-modal dyadic corpus used in [15, 16] that employs unrehearsed story-telling to elicit a controlled comparison of listener behavior in dyadic rapport across three language/cultural groups: American English (recorded in the United States), Mexican Spanish (recorded in the United States), and Iraqi Arabic (recorded in the United States and Amman, Jordan). Each pair of individuals was audio- and video-recorded performing their assigned task. All of these dyads were close acquaintances or family members with assumed well-established rapport from the same language/cultural group. One of them played the "speaker" role, and the other played the "listener" role. The "speaker" participant viewed the six minute "Pear Film", developed in [23] for language-independent elicitation. Afterward, the speaker related the story to the active and engaged listener, who would need to retell the story later based only on the information they obtained from the speaker.

All recordings have been fully transcribed and time-aligned to the audio, using a semi-automated procedure. An initial, coarse manual transcription at the phrase level, according to the silence- (non-speech-) delimited intervals, was converted to a full word and phone alignment using CUSonic [24], applying its language porting functionality to Spanish and Arabic.

In addition to transcription and alignment, the corpus is being annotated for other multi-modal behaviors, including gestures, gaze, blink, vocal backchannel, and head nod.

2.1. Annotation

The experiments in this paper use a subset of the corpus that has been fully transcribed, aligned, and annotated for vocal turn and backchannel and headnod. The set comprises 13 Arabic ($\approx$ 1 hour total), 15 English ($\approx$ 1.5 hours total), and 15 Spanish ($\approx$ 1 hour total) dyads approximately balanced for gender and speaker role, each completing a single narrative retelling task.

Vocal backchannel All spans of speaker and listener speech were annotated with one of three tags:

- Speaking turn: the individual holds the floor
- Vocal backchannel: the individual speaks, indicating continued attention, but does not take the floor
- Sentence completion: the individual performs a collaborative completion of the other's utterance, but does not take the floor.

For analysis, we group together the categories where the user does not take the floor and distinguish these as "vocal backchannel" in contrast to speaking turn. This grouping reflects the very low rate of the sentence completion class.

Head nod All spans of speaker and listener head nodding were annotated. Continuous periods of nodding were annotated as a single span, rather than annotating each oscillation.

Feature Type	Description
Pitch	Mean over last 300ms
	Slope over last 300ms
Intensity	Mean over last 300ms
Voice quality	NHR over last 1000ms
Speaking rate	Words/sec over IPU
Duration	Words over IPU

Table 1: Prosodic features for analysis

3. Analysis Conditions

Here we focus on the listener side of our quasi-monologic story-telling interaction to identify those points at which the listener produces a verbal backchannel or head nod or takes a speaking turn. Consistent with [14][1], we take as our unit of analysis an interpausal unit (referred to as IPU or utterance), a contiguous span of speech by a speaker bounded by at least 50 ms of silence or non-speech, such as breath sounds, according to our transcription and alignment. For consistency and comparability in these experiments, we consider only those listener behaviors - turns, verbal backchannels, and nods - that begin in the silence interval between two speaker-side utterances. All feature extraction intervals are then relative to the end of the immediately preceding speaker-side utterance. We compare those speaker utterances that precede and potentially cue these specific listener behaviors to those that do not. To allow our cross-language/cultural comparisons, we perform similar analyses independently for each group and contrast our findings.

3.1. Feature Extraction

Similar to [14], we extracted pitch, intensity, voice quality, speaking rate, and utterance duration features for each speaker-channel utterance. The full list is in Table 1. Pitch, intensity, and voice quality features are extracted using Praat's [25] "To Pitch...", "To Intensity", and "To Harmonicity...", respectively. Measures are computed for 300ms or 1000ms intervals calculated from the end of the utterance; however, voice quality measures were calculated only over the aligned vowel spans within the intervals. All durational, word, and phone position information is obtained from the semi-automatic alignment described above. For all acoustic-prosodic features, we performed a log-scaled, z-score normalization per speaker/dyad before analysis.

4. Analysis of Vocal Backchannels

For each of the measures above, we compare the speaker-side utterances immediately preceding listener vocal backchannels (BC) to those preceding listener speaking turns (T) and to those with no listener verbal activity (N)[2]. In each case, we perform Kruskal-Wallis tests, followed by Tukey post-hoc tests as necessary, to identify significant differences on these measures across these different listener verbal conditions; differences are considered significant at the $p < 0.05$ level. The proportion of speaker utterances immediately preceding listener verbal backchannels (BC) and speaking turns (T) for each language in the corpus is

[1] Due to the relatively small absolute number of listener backchannels and the difficulty of POS tagging dialectal Arabic and conversational Spanish speech, we excluded POS tag sequence analysis. We also did not have the resources for manual ToBI intonation annotation

[2] These results are a significant refinement of [15, 16] by breaking down the analysis by type of listener verbal behavior.

	BC	T		BC	T		BC	T
A	19%	6.7%	E	4.6%	3.8%	S	4.4%	2.1%

Table 2: Percentage of IPUs followed by listener turns (T) or vocal backchannels (BC), by type and language: (A)rabic, (E)nglish, and (S)panish

Arabic	12.4%	English	6.4%	Spanish	6.7%

Table 4: Rates of IPUs followed by listener head nod, by language.

shown in Table 2. We find significant effects for many of these measures across the different languages. Key differences appear in Table 3.

The overall trends are as follows. Across all three language/cultural groups, verbal backchannels are associated with lower speaking rate, longer utterance duration, and lower voice quality measures on preceding speaker utterances than contexts with no listener verbalizations.

Verbal backchannels are associated with significantly lower pitch mean than are N cases in English and Arabic. Both English and Spanish exhibit lower intensity for BC than N; however, verbal backchannels in Arabic do not. The characterization of cues prior to listener turns is more complex. English listener turns are preceded by significantly higher mean pitch than are BC cases, while Arabic turn cases exhibit lower pitch than N, though no lower than BC. Differences in pitch slope are not significant for any group.

Overall, the presence of lowness, in pitch and/or intensity, as a cue for vocal backchannel across these language/cultural groups is consistent with prior findings by [5, 6, 7], though it contrasts with findings of [14] which linked increases in pitch and intensity with back-channels in English task-oriented dialog.

5. Head Nod Analysis

Analogously to the discussion of listener verbal backchannel and turn above, we compared those speaker utterances immediately preceding listener head nods (HN) to those with no listener nod (N); the proportion of HN in the corpus appear in Table 4. A summary of results appears in Table 5.

Curiously, although verbal backchannels and head nods are frequently viewed as comparable backchannels, in our corpus they are associated with quite different prosodic cues in the preceding speaker utterances. In many cases, there is no reliably significant prosodic difference between utterances preceding a listener head nod and those without.

Verbal backchannels are associated with low pitch and low intensity in English, whereas head nods appear to be linked to high pitch in the preceding speaker utterance in both English and Spanish. While similar voice quality contrasts appear in English and Spanish for verbal backchannels and head nods, only English maintains its speaking rate and duration contrasts across both verbal and non-verbal backchannels. For Arabic, head nods are associated with no significant prosodic differences, while pitch plays an important role in the verbal case. This divergence is particularly surprising given that approximately 50% of verbal backchannels cooccur with headnods overall; percentages of IPUs with feedback overlap are shown in Table 6. In the case of many Arabic female listeners, all verbal backchannels are produced within one second of a headnod.

	HNBC	HNT		HNBC	HNT		HNBC	HNT
A	4.2%	1.2%	E	1.6%	0.8%	S	1.2%	0.5%

Table 6: Rates of IPUs with joint listener verbal feedback and head nod, by type and language.

Language	Intensity	Pitch
Arabic	$N >> T$	$N >> BC, T$; $HN >> BC$
English	$N >> BC >> T$; $HN >> T$	$HN, T >> N, BC$
Spanish	$N >> T$	$HN >> N, BC$

Table 7: Contrasting cues for listener verbal backchannel (BC) including those with concurrent head, turn (T), isolated head nod (HN), and None (N).

6. Joint Analysis of Verbal Backchannel and Head Nod Cues

We will consider these two different types of social resonance behavior - verbal backchannel and non-verbal head nod - together to better understand this apparent conflict. The apparent divergence in cues to verbal and non-verbal backchannels is most confusing in the case of concurrent verbal backchannels and head nods. We separate out the IPUs preceding these joint verbal and non-verbal backchannels (HN-BC) and compare them to the utterances that precede only verbal backchannels (BC) and those which precede only head nods (HN).

[22] provides an analysis of backchannel, including head nod function from less marked continuers to convergence tokens, engaged response tokens, and information receipt tokens. Building on that analysis, [18] identifies differences in backchannel form associated with these functions. In particular, the continuum is characterized by increases in duration and head nod magnitude, with continuers having the lowest duration and magnitude. To assess whether the isolated head nods and those in conjunction with backchannels might represent such different classes, we compare the durations of the head nods in the HN-BC and HN conditions[3]. By Kruskal-Wallis test, we find that nod durations for HN are significantly longer than those for HN-BC for all three language/cultural groups. This contrast argues for treating these nod groups as two different classes.

Thus we group the utterances that precede joint verbal and non-verbal backchannels into the verbal backchannel (BC) class, keeping those that precede only headnods (HN) separate. These two categories exhibit significant differences in pitch across all three languages. The pure HN class is also significantly higher in pitch than the class of utterances that do not precede either verbal or non-verbal backchannels, for both English and Spanish. Detailed results for pitch and intensity are found in Table 7.

7. Discussion and Conclusions

We have highlighted similarities and differences in use of a range of prosodic cues to backchannel behaviors, including pitch, intensity, voice quality, duration, and speaking rate across three different language/cultural groups. In particular, we have

[3]Due to very low frequency of the conjoint turn and head nod class, we restrict the statistical analysis to the HN-BC class.

Language	Intensity Mean	Pitch Mean	Spkg Rate	Duration	NHR
Arabic	$N, BC >> T$	$N >> BC, T$	$N, BC >> T$	$BC >> N >> T$	$N >> BC$
English	$N >> BC >> T$	$N >> BC; T >> BC$	$N >> BC$	$BC >> N$	$N >> BC$
Spanish	$N >> T$	n.s.	$N >> BC$	$BC >> N$	$N >> BC$

Table 3: Significant differences in prosodic features for utterances preceding listener verbal backchannels (BC), turns (T), or no listener behavior (N). $X >> Y$ indicates X significantly greater than Y.

Language	Intensity Mean	Pitch Mean	Spkg Rate	Duration	NHR
Arabic	n.s.	n.s.	n.s.	n.s.	n.s.
English	n.s.	$HN >> N$	$N >> HN$	$HN >> N$	$N >> HN$
Spanish	$N >> HN$	$HN >> N$	n.s.	n.s.	$N >> HN$

Table 5: Significant differences in prosodic features for utterances preceding listener head nods (HN) or no listener behavior (N). $X >> Y$ indicates X significantly greater than Y; n.s. indicates differences do not reach significance.

identified significant differences in prosodic cues, especially pitch, between verbal backchannels and head nods attested across languages. These contrasts argue that, far from being a monolithic class, these verbal and non-verbal backchannel behaviors are produced in response to substantially different cues. Furthermore, head nods themselves appear to fall into potentially two different categories, one appearing and patterning with verbal backchannels and another contrasting with them in prosodic measures, such as pitch.

We plan to further investigate whether and how these contrasts in prosodic cues reflect differences in function or form of these social resonance behaviors. We further hope to exploit these cues and insights into their cross-cultural similarities and differences to improve automatic prediction of backchannels.

8. Acknowledgments

This work was supported by NSF BCS #0725919. We would like to thank the team of annotator/analysts and especially Frederica Lippman and Edward King for the alignments.

9. References

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

[2] T. Novinger, *Intercultural Communication: A Practical Guide*. Austin, TX: University of Texas Press, 2001.

[3] D. Matsumoto, S. H. Yoo, S. Hirayama, and G. Petrova, "Validation of an individual-level measure of display rules: The display rule assessment inventory (DRAI)," *Emotion*, vol. 5, pp. 23–40, 2005.

[4] O. M. Watson, *Proxemic Behavior: A Cross-cultural Study*. The Hague: Mouton, 1970.

[5] 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.

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

[7] 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.

[8] A. Kendon, *Gesture: Visible Action as Utterance*. Cambridge University Press, 2004.

[9] L. Tickle-Degnen and R. Rosenthal, "The nature of rapport and its nonverbal correlates," *Psychological Inquiry*, vol. 1, no. 4, pp. 285–293, 1990.

[10] 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.

[11] 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.

[12] 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.

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

[14] ——, "Turn-taking cues in task-oriented dialogue," *Computer Speech and Language*, vol. 25, no. 3, pp. 601–634, 2011.

[15] G.-A. Levow, S. Duncan, and E. King, "Cross-cultural investigation of prosody in verbal feedback in interactional rapport," in *Proceedings of Interspeech 2010*, 2010, pp. 286–289.

[16] S. Wang and G.-A. Levow, "Contrasting multi-lingual prosodic cues to predict verbal feedback for rapport," in *Proceedings of ACL-2011*, 2011, pp. 614–619.

[17] D. Knight, S. Bayoumi, S. Mills, A. Crabtree, S. Adolphs, T. Pridmore, and C. Carter, "Beyond the text: Construction and analysis of multi-modal linguistic corpora," in *Proceedings of the 2nd International Conference on e-Social Science*, 2006.

[18] D. Knight, "A multi-modal corpus approach to the analysis of backchanneling behaviour," Ph.D. dissertation, University of Nottingham, 2009.

[19] D. Herrera, D. Novick, D. Jan, and D. Traum, "The UTEP-ICT cross-cultural multiparty multimodal dialog corpus," in *MMC 2010*, 2010.

[20] L.-P. Morency, I. de Kok, and J. Gratch, "Predicting listener backchannels: A probabilistic multimodal approach," in *Proceedings of Conference on Intelligent Virutal Agents (IVA 2008)*, 2008.

[21] F. Khan, B. Mutlu, and X. Zhu, "Modeling social cues: Effective features for predicting listener nods," in *Proceedings of the NIPS Workshop on Modeling Human Communication Dynamics*, 2010.

[22] A. O'Keeffe and S. Adolphs, *Using a corpus to look at variational pragmatics: Response tokens in British and Irish discourse*. John Benjamins, 2008, pp. 69–98.

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

[24] 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.

[25] P. Boersma and D. Weenink, "Praat: doing phonetics by computer [computer program]. version 5.2.26," 2011, http://www.praat.org.