Contrasting Multi-Lingual Prosodic Cues to Predict Verbal Feedback for Rapport

Siwei Wang
Department of Psychology
University of Chicago
Chicago, IL 60637 USA
siweiw@cs.uchicago.edu

Gina-Anne Levow
Department of Linguistics
University of Washington
Seattle, WA 98195 USA
levow@uw.edu

Abstract

Verbal feedback is an important information source in establishing interactional rapport. However, predicting verbal feedback across languages is challenging due to language-specific differences, inter-speaker variation, and the relative sparseness and optionality of verbal feedback. In this paper, we employ an approach combining classifier weighting and SMOTE algorithm oversampling to improve verbal feedback prediction in Arabic, English, and Spanish dyadic conversations. This approach improves the prediction of verbal feedback, up to 6-fold, while maintaining a high overall accuracy. Analyzing highly weighted features highlights widespread use of pitch, with more varied use of intensity and duration.

1 Introduction

Culture-specific aspects of speech and nonverbal behavior enable creation and maintenance of a sense of rapport. Rapport is important because it is known to enhance goal-directed interactions and also to promote learning. Previous work has identified cross-cultural differences in a variety of behaviors, for example, nodding (Maynard, 1990), facial expression (Matsumoto et al., 2005), gaze (Watson, 1970), cues to vocal back-channel (Ward and Tsukuhara, 2000; Ward and Al Bayyari, 2007; Rivera and Ward, 2007), nonverbal back-channel (Bertrand et al., 2007)), and coverbal gesturing (Kendon, 2004).

Here we focus on the automatic prediction of listener verbal feedback in dyadic unrehearsed storytelling to elucidate the similarities and differences in three language/cultural groups: Iraqi Arabic-, Mexican Spanish-, and American English-speaking cultures. (Tickle-Degnen and Rosenthal, 1990) identified coordination, along with positive emotion and mutual attention, as a key element of interactional rapport. In the verbal channel, this coordination manifests in the timing of contributions from the conversational participants, through turn-taking and back-channels. (Duncan, 1972) proposed an analysis of turn-taking as rule-governed, supported by a range of prosodic and non-verbal cues. Several computational approaches have investigated prosodic and verbal cues to these phenomena. (Shriberg et al., 2001) found that prosodic cues could aid in the identification of jump-in points in multi-party meetings. (Cathcart et al., 2003) employed features such as pause duration and part-of-speech (POS) tag sequences for back-channel prediction. (Gravano and Hirschberg, 2009) investigated back-channel-inviting cues in task-oriented dialog, identifying increases in pitch and intensity as well as certain POS patterns as key contributors. In multi-lingual comparisons, (Ward and Tsukuhara, 2000; Ward and Al Bayyari, 2007; Rivera and Ward, 2007) found pitch patterns, including periods of low pitch or drops in pitch, to be associated with eliciting back-channels across Japanese, English, Arabic, and Spanish. (Herrera et al., 2010) collected a corpus of multi-party interactions among American English, Mexican Spanish, and Arabic speakers to investigate cross-cultural differences in proxemics, gaze, and turn-taking. (Levow et al., 2010) identified contrasts in narrative length and rate of verbal feedback in recordings of American English-, Mexi-

Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics:shortpapers, pages 614–619,
Portland, Oregon, June 19-24, 2011.

can Spanish-, and Iraqi Arabic-speaking dyads. This work also identified reductions in pitch and intensity associated with instances of verbal feedback as common, but not uniform, across these groups.

2 Multi-modal Rapport Corpus

To enable a more controlled comparison of listener behavior, we collected a multi-modal dyadic corpus of unrehearsed story-telling. We audio- and video-recorded pairs of individuals who were close acquaintances or family members with, we assumed, well-established rapport. One participant viewed a six minute film, the “Pear Film” (Chafe, 1975), developed for language-independent elicitation. In the role of Speaker, this participant then related the story to the active and engaged Listener, who understood that they would need to retell the story themselves later. We have collected 114 elicitations: 45 Arabic, 32 Mexican Spanish, and 37 American English.

All recordings have been fully transcribed and time-aligned to the audio using a semi-automated procedure. We convert an initial manual coarse transcription at the phrase level to a full word and phone alignment using CUSonic (Pellom et al., 2001), applying its language porting functionality to Spanish and Arabic. In addition, word and phrase level English glosses were manually created for the Spanish and Arabic data. Manual annotation of a broad range of nonverbal cues, including gaze, blink, head nod and tilt, fidget, and coverbal gestures, is underway. For the experiments presented in the remainder of this paper, we employ a set of 45 vetted dyads, 15 in each language.

Analysis of cross-cultural differences in narrative length, rate of listener verbal contributions, and the use of pitch and intensity in eliciting listener vocalizations appears in (Levow et al., 2010). That work found that the American English-speaking dyads produced significantly longer narratives than the other language/cultural groups, while Arabic listeners provided a significantly higher rate of verbal contributions than those in the other groups. Finally, all three groups exhibited significantly lower speaker pitch preceding listener verbal feedback than in other contexts, while only English and Spanish exhibited significant reductions in intensity. The current paper aims to extend and enhance these findings by exploring automatic recognition of speaker prosodic contexts associated with listener verbal feedback.

3 Challenges in Predicting Verbal Feedback

Predicting verbal feedback in dyadic rapport in diverse language/cultural groups presents a number of challenges. In addition to the cross-linguistic, cross-cultural differences which are the focus of our study, it is also clear that there are substantial inter-speaker differences in verbal feedback, both in frequency and, we expect, in signalling. Furthermore, while the rate of verbal feedback differs across language and speaker, it is, overall, a relatively infrequent phenomenon, occurring in as little as zero percent of pausal intervals for some dyads and only at an average of 13-30% of pausal intervals across the three languages. As a result, the substantial class imbalance and relative sparsity of listener verbal feedback present challenges for data-driven machine learning methods. Finally, as prior researchers have observed, provision of verbal feedback can be viewed as optional. The presence of feedback, we assume, indicates the presence of a suitable context; the absence of feedback, however, does not guarantee that feedback would have been inappropriate, only that the conversant did not provide it.

We address each of these issues in our experimental process. We employ a leave-one-dyad-out cross-validation framework that allows us to determine overall accuracy while highlighting the different characteristics of the dyads. We employ and evaluate both an oversampling technique (Chawla et al., 2002) and class weighting to compensate for class imbalance. Finally, we tune our classification for the recognition of the feedback class.

4 Experimental Setting

We define a Speaker pausal region as an interval in the Speaker’s channel annotated with a contiguous span of silence and/or non-speech sounds. These Speaker pausal regions are tagged as ’Feedback (FB)’ if the participant in the Listener role initiates verbal feedback during that interval and as ’No Feedback (NoFB)’ if the Listener does not. We aim to characterize and automatically classify each such

Arabic	English	Spanish
0.30 (0.21)	0.152 (0.10)	0.136 (0.12)

Table 1: Mean and standard deviation of proportion of pausal regions associated with listener verbal feedback

region. We group the dyads by language/cultural group to contrast the prosodic characteristics of the speech that elicit listener feedback and to assess the effectiveness of these prosodic cues for classification. The proportion of regions with listener feedback for each language appears in Table 1.

4.1 Feature Extraction

For each Speaker pausal region, we extract features from the Speaker's words immediately preceding and following the non-speech interval, as well as computing differences between some of these measures. We extract a set of 39 prosodic features motivated by (Shriberg et al., 2001), using Praat's (Boersma, 2001) "To Pitch..." and "To Intensity...". All durational measures and word positions are based on the semi-automatic alignment described above. All measures are log-scaled and z-score normalized per speaker. The full feature set appears in Table 2.

4.2 Classification and Analysis

For classification, we employ Support Vector Machines (SVM), using the LibSVM implementation (C-C.Cheng and Lin, 2001) with an RBF kernel. For each language/cultural group, we perform 'leave-one-dyad-out' cross-validation based on F-measure as implemented in that toolkit. For each fold, training on 14 dyads and testing on the last, we determine not only accuracy but also the weight-based ranking of each feature described above.

Managing Class Imbalance Since listener verbal feedback occurs in only 14-30% of candidate positions, classification often predicts only the majority 'NoFB' class. To compensate for this imbalance, we apply two strategies: reweighting and oversampling. We explore increasing the weight on the minority class in the classifier by a factor of two or four. We also apply SMOTE (Chawla et al., 2002) oversampling to double or quadruple the number of minority class training instances. SMOTE oversampling creates new synthetic minority class instances by identifying $k = 3$ nearest neighbors and inducing a new instance by taking the difference between a sample and its neighbor, multiplying by a factor between 0 and 1, and adding that value to the original instance.

5 Results

Table 4 presents the classification accuracy for distinguishing FB and NoFB contexts. We present the overall class distribution for each language. We then contrast the minority FB class and overall accuracy under each of three weighting and oversampling settings. The second row has no weighting or oversampling; the third has no weighting with quadruple oversampling on all folds, a setting in which the largest number of Arabic dyads achieves their best performance. The last row indicates the oracle performance when the best weighting and oversampling setting is chosen for each fold.

We find that the use of reweighting and oversampling dramatically improves the recognition of the minority class, with only small reductions in overall accuracy of 3-7%. Under a uniform setting of quadruple oversampling and no reweighting, the number of correctly recognized Arabic and English FB samples nearly triples, while the number of Spanish FB samples doubles. We further see that if we can dynamically select the optimal training settings, we can achieve even greater improvements. Here the number of correctly recognized FB examples increases between 3- (Spanish) and 6-fold (Arabic) with only a reduction of 1-4% in overall accuracy. These accuracy levels correspond to recognizing between 38% (English, Spanish) and 73% (Arabic) of the FB instances. Even under these tuned conditions, the sparseness and variability of the English and Spanish data continue to present challenges.

Finally, Table 3 illustrates the impact of the full range of reweighting and oversampling conditions. Each cell indicates the number of folds in each of Arabic, English, and Spanish respectively, for which that training condition yields the highest accuracy. We can see that the different dyads achieve optimal results under a wide range of training conditions.

Feature Type	Description	Feature IDs
Pitch	5 uniform points across word	pre_0,pre_0.25,pre_0.5,pre_0.75,pre_1
		post_0,post_0.25,post_0.5,post_0.75,post_1
	Maximum, minimum, mean	pre_pmax, pre_pmin, pre_pmean
		post_pmax, post_pmin, post_pmean
	Differences in max, min, mean	diff_pmax, diff_pmin, diff_pmean
	Difference b/t boundaries	diff_pitch_endbeg
	Start and end slope	pre_bslope, pre_eslope, post_bslope, post_eslope
	Difference b/t slopes	diff_slope_endbeg
Intensity	Maximum, minimum, mean	pre_imax, pre_imin, pre_imean
		post_imax,post_imin, post_imean
	Difference in maxima	diff_imax
Duration	Last rhyme, last vowel, pause	pre_rdur, pre_vdur, post_rdur, post_vdur, pause_dur
Voice Quality	Doubling & halving	pre_doub, pre_half,post_doub,post_half

Table 2: Prosodic features for classification and analysis. Features tagged 'pre' are extracted from the word immediately preceding the Speaker pausal region; those tagged 'post' are extracted from the word immediatey following.

weight	1	2	4
no SMOTE	1,2,3	2,2,2	1,0,3
SMOTE Double	1,0,2	1,2,0	2,2,1
SMOTE Quad	3,0,0	1,2,2	3,6,2

Table 3: Varying SVM weight and SMOTE ratio. Each cell shows # dyads in each language (Arabic, English, Spanish) with their best performance with this setting.

	Arabic	English	Spanish
Overall	478 (1405)	395 (2659)	173 (1226)
Baseline	53 (950)	23 (2167)	23 (1066)
S=2, W=1	145 (878)	67 (2120)	47 (1023)
Oracle	347 (918)	152 (2033)	68 (1059)

Table 4: Row 1: Class distribution: # FB instances (# total instances). Rows 2-4: Recognition under different settings: # FB correctly recognized (total # correct)

6 Discussion: Feature Analysis

To investigate the cross-language variation in speaker cues eliciting listener verbal feedback, we conduct a feature analysis. Table 5 presents the features with highest average weight for each language assigned by the classifier across folds, as well as those distinctive features highly ranked for only one language.

We find that the Arabic dyads make extensive and distinctive use of pitch in cuing verbal feedback, from both preceding and following words, while placing little weight on other feature types. In contrast, both English and Spanish dyads exploit both pitch and intensity features from surrounding words. Spanish alone makes significant use of both vocalic and pause duration. We also observe that, although there is substantial variation in feature ranking across speakers, the highly ranked features are robustly employed across almost all folds.

7 Conclusion

Because of the size of our data set, it may be premature to draw firm conclusion about differences between these three language groups based on this analysis. The SVM weighting and SMOTE oversampling strategy discussed here is promising for improving recognition on imbalanced class data. This strategy substantially improves the prediction

Most Important Features		
Arabic	English	Spanish
pre_pmax	pre_pmean	pre_min
pre_pmean	post_pmean	post_0.5
pre_0.25	post_0.5	post_0.75
pre_0.5	post_0.75	post_1
pre_0.75	post_1	pre_imax
pre_1	diff_pmin	pre_imean
post_pmin	pre_imax	post_imax
post_bslope	pre_imean	pause_dur
diff_pmin	post_imean	pre_vdur

Most Distinctive Features		
Arabic	English	Spanish
post_pmin	post_pmean	post_0
post_bslope	post_0.25	post_eslope
pre_0.25		pre_eslope
pre_0.5		post_vdur
pre_1		pre_imean

Table 5: Highest ranked and distinctive features for each language/cultural group

of verbal feedback. The resulting feature ranking also provides insight into the contrasts in the use of prosodic cues among these language cultural groups, while highlighting the widespread, robust use of pitch features.

In future research, we would like to extend our work to exploit sequential learning frameworks to predict verbal feedback. We also plan to explore the fusion of multi-modal features to enhance recognition and increase our understanding of multi-modal rapport behavior. We will also work to analyze how quickly people can establish rapport, as the short duration of our Spanish dyads poses substantial challenges.

8 Acknowledgments

We would like to thank our team of annotator/analysts for their efforts in creating this corpus, and Danial Parvaz for the development of the Arabic transliteration tool. We are grateful for the insights of Susan Duncan, David McNeill, and Dan Loehr. This work was supported by NSF BCS#: 0729515. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

References

R. Bertrand, G. Ferre, P. Blache, R. Espesser, and S. Rauzy. 2007. Backchannels revisited from a multimodal perspective. In *Auditory-visual Speech Processing*, The Netherlands. Hilvarenbeek.

P. Boersma. 2001. Praat, a system for doing phonetics by computer. *Glot International*, 5(9–10):341–345.

C-C.Cheng and C-J. Lin. 2001. LIBSVM:a library for support vector machines. Software available at: http://www.csie.ntu.edu.tw/ cjlin/libsvm.

N. Cathcart, J. Carletta, and E. Klein. 2003. 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*, pages 51–58.

W. Chafe. 1975. The Pear Film.

Nitesh Chawla, Kevin Bowyer, Lawrence O. Hall, and W. Philip Legelmeyer. 2002. SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16:321–357.

S. Duncan. 1972. Some signals and rules for taking speaking turns in conversations. *Journal of Personality and Social Psychology*, 23(2):283–292.

A. Gravano and J. Hirschberg. 2009. Backchannel-inviting cues in task-oriented dialogue. In *Proceedings of Interspeech 2009*, pages 1019–1022.

David Herrera, David Novick, Dusan Jan, and David Traum. 2010. The UTEP-ICT cross-cultural multiparty multimodal dialog corpus. In *Proceedings of the Multimodal Corpora Workshop: Advances in Capturing, Coding and Analyzing Multimodality (MMC 2010)*.

A. Kendon. 2004. *Gesture: Visible Action as Utterance*. Cambridge University Press.

G.-A. Levow, S. Duncan, and E. King. 2010. Cross-cultural investigation of prosody in verbal feedback in interactional rapport. In *Proceedings of Interspeech 2010*.

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

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

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

A. Rivera and N. Ward. 2007. Three prosodic features that cue back-channel in Northern Mexican Spanish. Technical Report UTEP-CS-07-12, University of Texas, El Paso.

E. Shriberg, A. Stolcke, and D. Baron. 2001. Can prosody aid the automatic processing of multi-party meetings? evidence from predicting punctuation, disfluencies, and overlapping speech. In *Proc. of ISCA Tutorial and Research Workshop on Prosody in Speech Recognition and Understanding*.

Linda Tickle-Degnen and Robert Rosenthal. 1990. The nature of rapport and its nonverbal correlates. *Psychological Inquiry*, 1(4):285–293.

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

N. Ward and W. Tsukuhara. 2000. Prosodic features which cue back-channel responses in English and Japanese. *Journal of Pragmatics*, 32(8):1177–1207.

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