EMPLOYING BOOSTING TO COMPARE CUES TO VERBAL FEEDBACK IN MULTI-LINGUAL DIALOG

Gina-Anne Levow *

University of Washington
Linguistics
Seattle, WA USA

Siwei Wang

Argonne National Laboratory
Mathematics and Computer Science
Argonne, IL USA

ABSTRACT

Verbal feedback provides important cues in establishing interactional rapport. The challenge of recognizing contexts for verbal feedback largely arises from relative sparseness and optionality. In addition, cross-language and inter-speaker variations can make recognition more difficult. In this paper, we show that boosting can improve accuracy in recognizing contexts for verbal feedback based on prosodic cues. In our experiments, we use dyads from three languages (English, Spanish and Arabic) to evaluate two boosting methods, generalized Adaboost and Gradient Boosting Trees, against Support Vector Machines (SVMs) and a naive baseline, with explicit oversampling on the minority verbal feedback instances. We find that both boosting methods outperform the baseline and SVM classifiers. Analysis of the feature weighting by the boosted classifiers highlights differences and similarities in the prosodic cues employed by members of these diverse language/cultural groups.

Index Terms— Spoken dialog, prosody, verbal feedback, boosting

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], nonverbal back-channel [8]), and coverbal gesturing [9] as well.

Here our focus is to develop an automatic classification framework that can successfully identify cues to listener verbal feedback in dyadic interactions involving unrehearsed story-telling. This classification will help to identify culture-specific factors, as well as similarities and differences in three language/cultural groups: Iraqi Arabic, Mexican Spanish and American English-speaking cultures. Furthermore, the cross-culture comparisons of our study will be employed to develop Listening Embodied Conversational Agents (ECAs), [10] that are able to produce culturally distinctive behaviors related to establishment and maintenance of rapport. Therefore, we employ an analysis-by-classification approach to identify prosodic cues produced by the speaker that can serve to signal appropriate times for listener vocalization.

[11] 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 back-channels among conversational participants. Foundational work by [12] established that conversational interaction is fundamentally rule-governed. Multi-modal cues including gaze, posture, nod and prosody were shown to cue turn-taking. In addition, [1, 13] 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 feedback in dyadic and multi-party scenarios based on a quantitative and computational perspective. In [14], prosodic cues were shown to be informative in identifying jump-in points in multi-party meetings. In [15], features from shallow processing, like pause duration and part-of-speech (POS) tag sequences, are shown to be helpful in predicting back-channels. In [16], it was reported that increases in pitch and intensity, as well as certain POS patterns, are key back-channel-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 back-channels in Japanese-, English-, Arabic- and Spanish-speakers. Recently, [17] investigated cultural differences in gaze, proxemics, and back-channel behavior in a multi-modal corpus of American English, Mexican Spanish, and Arabic speakers. [18] presented initial analyses of another multimodal 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 that all groups exploited reduced pitch, while only Spanish and English speakers employed reductions in intensity. Using the same corpus, [19] demonstrated improved prediction of listener verbal feedback through a combination of class reweighting and oversampling using Support Vector Machines trained on prosodic features.

2. MULTI-MODAL RAPPORT CORPUS

We employ the same multi-modal dyadic corpus used in [18, 19] that employs unrehearsed story-telling to elicit a controlled comparison of listener behavior in dyadic rapport across three language/cultural groups: American English, Mexican Spanish, and Iraqi Arabic. Each

*This work was supported by NSF BCS #: 0725919.

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

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. 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 [20] for language-independent elicitation. Afterward, the speaker related the story to the active and engaged listener, who would need to retell the story later only based 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 [21], applying its language porting functionality to Spanish and Arabic.

The experiments in this paper use a subset of the corpus that contains 15 dyads from each language group and 45 dyads in total and has been fully processed and verified.

3. CHALLENGES IN VERBAL FEEDBACK

There are a number of challenges in identifying contexts for verbal feedback using the multi-lingual dyadic rapport corpus. First, cross-linguistic, cross-cultural differences lead to differences in signaling and expectation for verbal feedback. Second, there are substantial inter-speaker differences in verbal feedback. Third, verbal feedback is, overall, an infrequent phenomenon. As shown in Table 1, verbal feedback occurs, on average, in 13% to 30% of all pause intervals, depending on the language. Furthermore, in some English- and Spanish-speaking dyads, listeners produce no instances of verbal feedback at all. As a result, the substantial class imbalance and relative sparsity of listener verbal feedback present challenges for data-driven machine learning methods, especially those that focus on empirical risk minimization. Finally, 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.

4. ADABOOST AND GRADIENT BOOSTING

Class imbalance is the major motivation that leads us to investigate boosting. The key idea of Adaboost, placing greater weight on misclassified samples, can potentially help the classifier to focus on recognizing samples of the minority class. This is because those samples tend to be misclassified, as the early stages of training are biased toward recognizing the majority class instances to obtain high overall accuracy. We will introduce two different boosting methods for classifying contexts for verbal feedback in the rest of this section. In our experiments, we employ an ensemble of 100 trees in each boosting approach. Each of the trees has $depth = 3$. We will discuss the performance in detail in Sections 6 and 7.

4.1. A Generalized Version of Adaboost on Decision Trees

Generalized Adaboost ("Adaboost real") was introduced in [22]. It is different from the original Adaboost in that, first, it allows weak hypotheses to have real-valued output, rather than output only in the restricted range $[-1, +1]$, and, second, it leaves the weighting of weak learners α_t open and allows flexibility for various tuning strategies. The basic steps of the generalized Adaboost algorithm are:

Given: $(x_1, y_1), ..., (x_m, y_m)$, where x_i is the i^{th} input feature vector, y_i is the corresponding i^{th} output label.

Initialize a weight distribution over all input samples $D_1(i) = 1/m$, such that every sample has the same weight.

For iterations $t = 1, ..., T$,

- Train weak learner using distribution D_t.
- Get weak hypothesis $h_t : X \rightarrow R$.
- Choose $\alpha_t \in R$.
- Update:

$$D_{t+1}(i) = \frac{D_t(i) exp(\alpha_t y_i h_t(x_i))}{Z_t} \tag{1}$$

  where Z_t is a normalization factor (chosen so that D_{t+1} is a distribution)

- Output the final hypothesis:

$$H(x) = sign(\sum_{t=1}^{T} \alpha_t h_t(x)) \tag{2}$$

4.2. Gradient Tree Boosting

Gradient Tree Boosting is a generalized boosting method that uses decision trees to perform optimization of arbitrary differentiable loss functions. This method was first introduced in [23] as an approach for prediction problems with continuous input space. Given training samples $(x_1, y_1), ..., (x_m, y_m)$, the goal of gradient boosting is to find an approximation $F(x)$ for every class that minimizes the loss function $L(y, F(x))$.

The basic steps of the algorithm are listed below:

Initialize $f_0(x) = argmin_\gamma \sum_{i=1}^{N} L(y_i, \gamma)$ in decision tree boost:

For iterations $t = 1$ to T : for every class y_i in Y:

- For $i = 1, 2, ..., N$ Compute

$$r_{it} = -[\frac{\partial L(y_i, f(x_i)))}{\partial f(x_i))}]_{f=f_{t-1}} \tag{3}$$

- Fit a regression tree to the targets r_{it} giving terminal regions $R_{jt}, j = 1, 2, ..., J_t$.
- For $j = 1, 2, ..., J_t$ compute the shrinkage γ_{jt}: the ratio of how much information the new tree should inherit from the old trees

$$\gamma_{jt} = argmin_\gamma \sum_{x_i \in R_{jt}} L(y_i, f_{t-1}(x_i) + \gamma) \tag{4}$$

- Update

$$f_t(x) = f_{t-1}(x) + \sum_{j=1}^{J_t} \gamma_{jt} I(x \in R_{jt}) \tag{5}$$

- Output $\hat{f}(x) = f_T(x)$

Compared to Adaboost, gradient boosting differs in two ways. First, every weak learner is optimized to reduce the error of recognizing a single class. For example, two decision trees are the minimum number we need in gradient boosting to perform two class classification, one for each class. Second, it only optimizes on weak learners, instead of weighting the samples differently as in Adaboost. Frequently employed loss functions $L(y, F)$ include the absolute error $|y - F|$ for regression and binomial log-likelihood for classification. In this paper, we employ binomial log-likelihood, the deviance loss $F_0(x) = \frac{1}{2} log \frac{1+\bar{y}}{1-\bar{y}}$, for our gradient boosting trees.

5. EXPERIMENTS

Every feature vector corresponds to a "speaker" pause region, a contiguous span of annotated silence and/or non-speech sounds in the channel of the participant in the "speaker" role. These pause regions are tagged as 'Feedback' if the listener initiates a verbal contribution during that interval and as 'No Feedback' otherwise. The focus of our experiments is to understand the characteristics of all pause regions and automatically classify them with respect to the occurrence of verbal feedback. In addition, we would also like to identify any variation between languages/cultures in using prosodic cues to signal verbal feedback. Therefore, we group our dyads by their language/cultural identities and perform the classification for each group separately. Table 1 shows the proportion of regions with verbal feedback in each language group.

5.1. Feature Extraction

We extracted pitch and energy features motivated by [24] around each pause region. The full list is in Table 2. All prosodic features are extracted from the words immediately preceding and following the non-speech interval, including the differences between some of these measurements. Both pitch features and energy features are extracted using Praat's [25] "To Pitch..." and "To Intensity". All durational and word position information is obtained from the semi-automatic alignment described above. For all features, we performed a log-scaled, z-score normalization per speaker/dyad before classification.

All measures are computed from the "speaker" channel. Pause duration is thus the duration of the interval of contiguous silence and/or non-speech sounds in the "speaker" channel, independent of listener behavior.

5.1.1. Prosodic Feature Analysis

We perform a one-way ANOVA analysis of our prosodic features for each language/cultural group to determine whether significant differences in these features are associated with verbal feedback and thus validate their suitability for classification. Table 3 lists those features which differ significantly between feedback and non-feedback instances, ranked by increasing p value.

We observe that the current pause duration is the most highly significant feature in cuing verbal feedback, across all three language/cultural groups. Pitch features, in particular, dominate the list of significantly different features in Arabic dyads. In contrast, both English and Spanish dyads show more significant differences in durational features and intensity features.

Most Important Features		
Arabic	English	Spanish
pause_dur**	pause_dur**	pause_dur**
post_pmin**	post_pmean**	post_imean**
pre_0.5 **	pre_imax**	pre_pausedur**
pre_0.75**	pre_pmin**	pre_imax**
pre_pmax**	post_imean**	pre_pmean**
pre_1**	pre_pausedur**	pre_bslope**
pre_0.25**	post_0.25**	post_0.5**
pre_pmean**	post_0.5**	post_0.75**
post_bslope**	post_0.75**	pre_pmin**
pre_bslope**	pre_pmax**	pre_0*
pre_0**	post_0**	pre_vdur*
post_imean**	pre_eslope**	post_bslope*
diff_pmin**	pre_1**	pre_imean*
pre_imax**	post_pmax**	post_vq*
pre_eslope*	post_pmin*	pre_0.25*
diff_slope_endbeg*	pre_0.75*	post_pmin*
pre_vdur*	post_bslope*	post_eslope*
diff_pitch_endbeg*	pre_vdur*	pre_eslope*
pre_rdur*	pre_imean*	diff_imax*
pre_pausedur*	post_imax*	pre_0.5*
post_eslope*	pre_rdur*	
post_0*		

Table 3. Highest ranked features for each language/cultural group by one-way ANOVA, *:$p <= 0.05$; **:$p <= 0.001$

5.2. Classification and Analysis Setting

We employ the Adaboost and Gradient Tree Boosting (GBT) implementations in the OpenCV package [1] to build two ensembles of 100 decision trees for each fold of leave-one-dyad out cross-validation. For each fold, we train on 14 dyads and test on the last.

5.3. Baseline Classifiers

We contrast the boosting methods with two baseline classifiers:

- Random Assignment: This is a naive baseline with prior knowledge about the average rate of verbal feedback in each language. We randomly label the pause intervals as 'feedback' or 'no feedback' based on the average rate for that group.
- Support Vector Machine (SVM): for comparison with [19], we use the LIBSVM [26] implementation with an RBF kernel to perform all SVM experiments.

5.4. Managing Class Imbalance

Considering that listener verbal feedback occurs only in 13%-30% of the candidate pause regions, classification is often biased to predict the majority 'no feedback' class. To further compensate for this imbalance, we apply SMOTE [27] oversampling to quadruple the number of minority verbal feedback class training instances, as in [19]. SMOTE oversampling synthesizes a new instance by identifying $k = 3$ nearest neighbors for every original instance and deriving features for new instances by taking the difference between a sample and its neighbor multiplied by a random factor between 0 and 1, and adding this value to the corresponding feature of the original instance.

[1] http://opencv.willowgarage.com/wiki/

Feature Type	Description	Feature IDs
Pitch	5 points uniformly sampled in voiced region of 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 of preceding and following words	pre_pmax, pre_pmin, pre_pmean post_pmax, post_pmin, post_pmean
	Differences in max, min, mean of preceding and following words	diff_pmax, diff_pmin, diff_pmean
	Difference between boundaries	diff_pitch_endbeg
	Start and end slope	pre_bslope, pre_eslope, post_bslope, post_eslope
	Difference between slopes	diff_slope_endbeg
Intensity	Maximum, minimum, mean of preceding and following words	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, pre_pausedur, pause_dur
Voice Quality	Doubling & halving by position	pre_vq_mid, pre_vq_end,post_vq_mid,post_vq_end

Table 2. Prosodic features for classification and analysis

6. RESULT

Table 4 presents the classification accuracy and F-measure for each class. We present results in three sections, corresponding to the three languages. In each section, the first four rows correspond to the results for random baseline, Support Vector Machine, Adaboost and GBT trained on the original dataset, respectively. The next three rows correspond to the three classifiers trained on the dataset with SMOTE-based quadrupling of minority class instances. In every row, the first column is accuracy; the first number corresponds to the accuracy of recognizing verbal feedback, and the numbers in parentheses are accuracy of the overall dataset. The last two numbers represent the F-measure for the non-feedback and feedback classes, respectively.

Both boosting approaches outperform the baselines in recognizing listener feedback. Adaboost yields 8% improvement in accuracy on the minority class in English dyads over the SVM baseline and 4% improvement in Spanish dyads over the naive baseline, with only a 2% drop in majority class F-measure. In English dyads, both boosting methods more than double the F-measure of listener feedback compared to the naive optimistic baselines.

After SMOTE oversampling, GBT improves the most in recognizing listener verbal feedback on Arabic and English dyads. Compared to its accuracy on the original dataset, GBT on SMOTE oversampled training instances improved 3% on English dyads and 5% on Arabic dyads. The overall accuracy also rises on English SMOTE oversampled training instances. This indicates that GBT achieves more balanced recognition using SMOTE oversampling on English data.

We can observe that the boosting approaches are able to achieve performance that can only be attained by SVM with explicit oversampling and data weighting. Using SMOTE oversampling, only in English dyads do both Adaboost and GBT achieve slightly more balanced recognition, but the F-measure did not improve much. In Spanish and Arabic dyads, SMOTE oversampling causes both boosting approaches to predict more listener feedback instances at the cost of overall accuracy. It seems that Adaboost and GBT are more successful than SVM in addressing recognition problems with imbalanced classes without additional data manipulation.

Arabic			
	% Accuracy	F_n	F_f
Random	31.2 (56.6)	0.67	0.34
SVM	21 (66)	0.77	0.31
Adaboost	38.9 (59.6)	0.69	0.41
GBT	37.7 (61.5)	0.71	0.41
SVM:S=2	39.1 (60.2)	0.70	0.41
Adab:S=2	39.5 (58.6)	0.68	0.41
GBT:S=2	43.1 (58.5)	0.68	0.43
English			
	% Accuracy	F_n	F_f
Random	12.2 (72.5)	0.84	0.12
SVM	18.4 (86.1)	0.92	0.29
Adaboost	26.6 (82.3)	0.90	0.32
GBT	27.1 (83.9)	0.91	0.35
SVM:S=2	18.4 (85)	0.92	0.28
Adab:S=2	31.1 (82.3)	0.90	0.36
GBT:S=2	30.9 (84.3)	0.91	0.39
Spanish			
	% Accuracy	F_n	F_f
Random	11 (75.2)	0.86	0.11
SVM	9.8 (86.8)	0.93	0.18
Adaboost	15.6 (83.1)	0.91	0.21
GBT	22.0 (83.4)	0.91	0.27
SVM:S=2	20.8 (79.1)	0.89	0.24
Adab:S=2	23.7 (81.0)	0.89	0.26
GBT:S=2	22.0 (82.2)	0.90	0.26

Table 4. Accuracy and F-measure for prediction of listener verbal feedback based on prosodic cues in three language/cultural groups. F_n: F-measure for pause intervals with no feedback, F_f: F-measure for pause intervals with verbal feedback

	Arabic	English	Spanish
Base	113(196,199)	76(107,109)	18(27,38)
S=2	137(199,217)	85(125,124)	22(41,38)

Table 7. Count of samples correctly recognized by both Adaboost and GBT, numbers in parentheses are the count of correctly recognized instances for (Adaboost,GBT) respectively

7. DISCUSSION: FEATURE ANALYSIS

Tables 5 and 6 present the 10 features with the highest average weighting for the ensemble of decision trees employed in Adaboost and GBT, respectively. Using this ranking data, we can investigate the following:

1. cross-language variations in prosodic cues eliciting listener verbal feedback,
2. differences in feature ranking under SMOTE oversampling, and
3. the difference between Adaboost and GBT in feature ranking.

As shown in Table 5, all three languages rank durational features as one of the most informative feature subsets. Using Adaboost, both English and Arabic dyads rely on the duration of the current pause and otherwise exclusively on pitch features to cue listener verbal feedback. Using GBT, the durational features of the words immediately preceding and following the pause regions are heavily employed, and the intensity features also help in improving classification. Spanish dyads also make significant use of both vocalic and pause duration, in both Adaboost and GBT implementations.

After we applied SMOTE oversampling, we found two main differences in the Adaboost tree ensemble:

1. The Adaboost tree ensemble makes more use of pitch difference features on Arabic dyads.
2. The voice quality features are employed in training the tree ensemble for English dyads.

Considering that GBT makes significant use of voice quality to train its tree ensemble for all three language/cultural groups, and that we obtained improved performance from Adaboost with oversampled training instances, it suggests that voice quality features are informative in eliciting listener verbal feedbacks.

The obviously different feature rankings between Adaboost and GBT motivate us to look into the dataset and compare how each sample is recognized by both classifiers. Table 7 presents the number of instances with listener feedback successfully recognized by both boosting classifiers. In all three languages, the overlap constitutes a significant portion in the original datasets, and the overlap portion rises in oversampling datasets. For those instances, it is possible that all three feature subsets of pitch, duration, and intensity will be able to provide important cues in recognizing listener verbal feedback.

8. CONCLUSION

The scale of our dataset makes it premature to present a firm argument about the differences between language groups and the differences between the feature rankings obtained by both boosting methods. However, it is clear that boosting shows relatively robust recognition performance without any additional SMOTE oversampling. Both boosting approaches outperform the SVM and the naive baseline consistently in all three language groups. In addition, the shallow trees we obtained from boosting give us insight into the similarities and differences in features exploited by different language/cultural groups. In future work we will investigate a sequential learning framework using feature induction, to better incorporate the temporal dynamics of rapport creation and maintenance.

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*, University of Texas Press, Austin, TX, 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*, Mouton, The Hague, 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," Tech. Rep. UTEP-CS-07-12, University of Texas, El Paso, 2007.

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

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

[10] J. Gratch, A. Okhmatovskaia, F. Lamothe, M. Marsella, R. Vander Werf, and L.-P. Morency, "Virtual rapport," in *6th International Conference on Intelligent Virtual Agents*, 2006.

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

[12] 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.

[13] 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.

[14] E. Shriberg, A. Stolcke, and D. Baron, "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*, 2001.

[15] 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.

Most Important Features					
Arabic adaboost	Arabic GBT	English adaboost	English GBT	Spanish adaboost	Spanish GBT
pause_dur	post_vq_end	pause_dur	post_vq_end	pause_dur	post_vq_end
pre_0.25	pre_pausedur	pre_pmean	pre_pausedur	pre_imax	pre_pausedur
pre_0.5	pre_vq_end	pre_pmax	pre_vq_end	pre_bslope	pre_vq_end
pre_vdur	pause_dur	pre_eslope	pause_dur	diff_pitch_endbeg	pause_dur
post_pmin	pre_rdur	post_pmin	pre_rdur	pre_vdur	pre_imax
post_pmax	post_rdur	pre_1	post_rdur	pre_1	post_rdur
pre_1	pre_imean	post_0.5	post_eslope	diff_pmean	pre_rdur
pre_pmax	diff_imax	post_0.75	pre_imax	post_0.5	pre_rdur
pre_bslope	pre_vdur	post_1	post_imax	post_pmax	diff_imax
pre_imax	post_imean	diff_slope_endbeg	pre_imean	diff_pmax	pre_vdur

Table 5. Highest ranked features for each language/cultural group and boosting algorithms

Most Important Features, Smote=2					
Arabic adaboost	Arabic GBT	English adaboost	English GBT	Spanish adaboost	Spanish GBT
pre_pausedur	post_vq_end	pause_dur	post_vq_end	post_vdur	post_vq_end
post_vq_mid	pre_pausedur	pre_pausedur	pre_pausedur	pre_imax	pre_pausedur
post_pmin	pre_vq_end	post_1	pre_vq_end	pre_vdur	pre_vq_end
pause_dur	pause_dur	pre_pmax	pause_dur	pre_rdur	pause_dur
diff_pmin	pre_rdur	pre_0.5	pre_rdur	pause_dur	pre_rdur
diff_imax	diff_pmin	diff_slope_endbeg	post_1	diff_imax	pre_imax
diff_pmax	post_rdur	post_0.75	post_rdur	pre_pmean	post_eslope
pre_vq_mid	pre_imean	post_vq_mid	diff_pmin	pre_bslope	diff_imax
post_pmax	diff_pmax	pre_vq_mid	post_imax	post_eslope	post_vdur
diff_slope_endbeg	pre_imax	post_pmax	diff_imax	pre_eslope	post_rdur

Table 6. Highest ranked features for each language/cultural group and boosting algorithms with Smote=2

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

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

[18] Gina-Anne Levow, Susan Duncan, and Edward King, "Cross-cultural investigation of prosody in verbal feedback in interactional rapport," in *Proceedings of Interspeech 2010*, 2010, pp. 286–289.

[19] Siwei Wang and Gina-Anne Levow, "Contrasting multi-lingual prosodic cues to predict verbal feedback for rapport," in *Proceedings of ACL-2011*, 2011, pp. 614–619.

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

[21] 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.

[22] Robert E. Schapire and Yoram Singer, "Improved boosting algorithms using confidence-rated predictions," in *Machine Learning*, 1999, pp. 80–91.

[23] Jerome H. Friedman, "Greedy function approximation: A gradient boosting machine," *Annals of Statistics*, 2000.

[24] Elizabeth Shriberg and Andreas Stolcke, "Prosody modeling for automatic speech recognition and understanding," *Mathematical Foundations of Speech and Language Processing*, vol. 138, pp. 105–114, 2004.

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

[26] C-C. Cheng and C-J. Lin, "LIBSVM: A library for support vector machines," 2001, Software available at: http://www.csie.ntu.edu.tw/ cjlin/libsvm.

[27] Nitesh Chawla, Kevin Bowyer, Lawrence O. Hall, and W. Philip Legelmeyer, "SMOTE: Synthetic minority oversampling technique," *Journal of Artificial Intelligence Research*, vol. 16, pp. 321–357, 2002.