Spoken Interaction Modeling for Automatic Assessment of Collaborative Learning

Jennifer Smith[1], *Harry Bratt*[1], *Colleen Richey*[1], *Nikoletta Bassiou*[1]
Elizabeth Shriberg[1], *Andreas Tsiartas*[1], *Cynthia D'Angelo*[2], *Nonye Alozie*[2]

[1]SRI International Speech Technology and Research (STAR) Laboratory
[2]SRI International Center for Technology in Learning (CTL)
{jennifer.smith, harry.bratt, colleen.richey, nikoletta.basiou,
elizabeth.shriberg, andreas.tsiartas, cynthia.dangelo, maggie.alozie} @sri.com

Abstract

Collaborative learning is a key skill for student success, but simultaneous monitoring of multiple small groups is untenable for teachers. This study investigates whether automatic audio-based monitoring of interactions can predict collaboration quality. Data consist of hand-labeled 30-second segments from audio recordings of students as they collaborated on solving math problems. Two types of features were explored: speech activity features, which were computed at the group level; and prosodic features (pitch, energy, durational, and voice quality patterns), which were computed at the speaker level. For both feature types, normalized and unnormalized versions were investigated; the latter facilitate real-time processing applications. Results using boosting classifiers, evaluated by F-measure and accuracy, reveal that (1) both speech activity and prosody features predict quality far beyond chance using majority-class approach; (2) speech activity features are the better predictors overall, but class performance using prosody shows potential synergies; and (3) it may not be necessary to session-normalize features by speaker. These novel results have impact for educational settings, where the approach could support teachers in the monitoring of group dynamics, diagnosis of issues, and development of pedagogical intervention plans.

Index Terms: speech activity detection, prosodic features, machine learning, student collaboration, collaborative learning, classroom education.

1. Introduction

Collaboration is an important skill for all students to learn and practice. It is an integral part of the push for competencies in 21st-century skills that students must master as they progress through school and into their careers [1]. Research has shown that students do not come to class with pre-existing knowledge on how to engage with their peers in collaborative activities and how best to work together productively in groups [2].

Management and assessment of collaborative learning tasks is difficult in typical classrooms when teachers need to monitor 10-15 groups with two to three students in each group [3]. Ideally, teachers would listen to peer interactions in each group for long enough to understand the progress of discourse, but very few teachers can do this well for so many groups. This project aims to build speech-based learning analytics for collaboration that could help teachers by identifying group processes and enabling teachers to target their interventions.

Although there are many approaches (e.g., keystroke data, written responses) for gathering diagnostic information about collaborative learning, most collaborative learning involves peer discourse. Automated analysis of peer discourse in collaborative learning has been successful [4, 5, 6], but almost all prior work depends on capturing discourse in a modality other than speech (e.g., student interaction in text-based chat rooms).

Speech data is uniquely central and authentic to peer discourse, but the field lacks key knowledge of automatically analyzed speech in small group collaboration. Some exploratory work has successfully developed speech analytics for a situation in which one student is asked to answer a question while on camera [7]. Other researchers have taken a different approach and are trying to apply speech analytics to very specific and sophisticated aspects of collaborative learning, such as "idea co-construction" [8] and "transactive contributions" [9]. This project focuses on simpler behaviors in collaborative situations.

This project investigates the feasibility and challenges of using the speech of small groups of students to determine the quality of each group's collaboration. We are developing feature detectors and using machine-learning techniques to find ways to aggregate the signal from these detectors to agree with the collaboration quality judgments of human observers. By analyzing features such as speaker, when each participant speaks, and how each participant speaks (e.g., rate of speech, loudness contours, intonation patterns), automatic systems can detect features of participation such as turn taking, crosstalk, emotion, and on-task behavior. This approach is tractable without the need to create a complete transcription. This paper presents findings from early analyses of the audio data.

2. Method

2.1. Data

Participating students worked together in groups of three on a set of collaborative math activities. The data set was collected during 86 collaborative sessions (about 15-20 minutes each). 141 middle school students (67 in sixth grade, 40 in seventh grade, and 34 in eighth grade) from six different schools participated in the study. The gender breakdown was evenly split across the students. Most students participated in 2 sessions with different group configurations. At the time of publication, annotation was completed for 43 of the 86 sessions; only those sessions were used in the following analyses.

The collaborative math activities included 12 separate problems, each of which required the three students to work together and talk to each other to coordinate their three answers to the problem. In addition to a video recording of each group, three

audio files were recorded from individual noise-cancelling microphones worn by students. These audio files were divided into segments that corresponded to the time the group spent on a particular math problem (items). The items were further divided into 30 second segments (windows). The last window of an item may be less than 30 seconds, depending on the length of the item. Approximately 25% of the windows were less than 30 seconds. Any windows less than 5 seconds long were discarded.

We used a subset of the corpus, comprised of 43 annotated sessions in which there were a total of 472 items and 1945 windows. The number of items per session varies (mean = 11.05, standard deviation = 3.69) because the groups of students completed a varying number of math problems and some groups went back to previously completed problems and worked on them again. Due to its dependence on both the number of items completed and item lengths, the number of windows per session varies even more (mean = 44.23, standard deviation = 10.28).

A team of five human judges, all education researchers, annotated the data at both the item and window level. All judges underwent training on the coding scheme and went through a calibration process to ensure reliability among them. After training, the average of the Cohen's kappa score for each pair of judges across four sessions was 0.612. During the annotation process, we selected additional calibration instances to ensure against significant drift on the application of the codes. All disagreements were discussed by the human judges and a final code was assigned.

Judges assigned one of four collaboration quality codes (Q codes) to both the items and the windows. The Q codes represented the degree to which the three students were collectively engaging in good collaboration. Importantly, the codes depend not on how much each student spoke, but whether and how much each student was engaged intellectually in the group problem solving. More successful collaboration occurs when students engage each others' thinking [10]. In other words, the collaboration quality codes differentiated between simple engagement (whether or not students were talking and paying attention) and intellectual engagement (whether or not the students were engaged in actively solving the problem at hand). The judges used both the audio and video recordings to make their decisions.

In descending order of collaboration quality, the four Q codes are defined as:

· Good Collaboration: All three students are working together and intellectually contributing to problem solving.

· Out in the Cold: Two students are working together, but the third is either not contributing or is being ignored.

· Follow the Leader: One student is taking the intellectual lead on solving the problem and is not bringing in others.

· Not Collaborating: No students are actively contributing to solving the problem; each is either off-task, or working independently.

Additionally, if students did not have an opportunity to collaborate (e.g., they were waiting for technical help), then the window was coded as not applicable (N/A). All items and windows coded as (N/A) were discarded from the data set. 435 items and 1623 windows remained. The Q codes are well distributed at both the item and window level, as seen in Table 1.

Table 1: *Distributions of the Q Codes.*

	Good	Cold	Follow	Not
Item Level	0.39	0.24	0.22	0.15
Window Level	0.32	0.28	0.22	0.18

2.2. Features

2.2.1. Speech activity features by group

To extract segmental and durational information from the speech signal, we used a Speech Activity Detection (SAD) system to identify the speech regions and exclude the silent and noisy regions. Our data collection and experimental setup allowed students to speak freely. This freedom resulted in audio recordings with overlapping speech from the three student participants.

To solve this problem, we used a SAD system based on speech variability [11, 12] and ran the system independently on each of the three student channels. Finally, we used a speech variability threshold optimized on a small set of four samples. The thresholded output on each audio channel was used to identify the student-specific speech signal and eliminate the noise, silence or cross-talk regions.

Features derived from SAD output capture information about the number, duration, and location of speech regions. They are thus related to features used in studies of dominance in multiparty meetings, [13, 14], although the features we used differ in details. We used the SAD output to create several duration-related statistics: total duration of speech for each student (Total Durations 1, 2 and 3); the duration in which each student was the only speaker (Solo Durations 1, 2 and 3); the duration in which each pair of students spoke simultaneously (Overlap Durations 1-2, 1-3 and 2-3); the duration in which all students spoke simultaneously (All Duration); and the duration in which all students were silent (No Duration).

Two of the SAD-derived statistics (All Duration and No Duration) characterize the group as a whole and were used directly as group-level features. The remaining sets of statistics (three each for Total, Solo and Overlap Durations) characterize SAD activity for individual speakers (or speaker pairs for Overlap Durations). To obtain group-level features for each set, we first converted the three statistics comprising each set to proportions (p_i) by dividing them by their sum. We then characterized the distribution of each set using Shannon entropy [15]:

$$\sum_{i=1}^{n} p_i \log_2 \left(\frac{1}{p_i} \right) \tag{1}$$

Here n is 3 as there are 3 speaker-level measurements per set. A maximum value ($\log_2 3 \approx 1.585$) indicates a window during which all three students (or overlapping pairs) are speaking equally, and a minimum value (0) indicates a window during which only one of the students (or overlapping pairs) speaks.

We used two versions of the SAD-derived features in analysis: raw values and group-normalized values. Features were normalized because speakers bring idiosyncratic SAD habits to the group. Normalized SAD features emphasize relative changes over the analysis windows; unnormalized features do not. We normalized by subtracting the mean of each feature across each session.

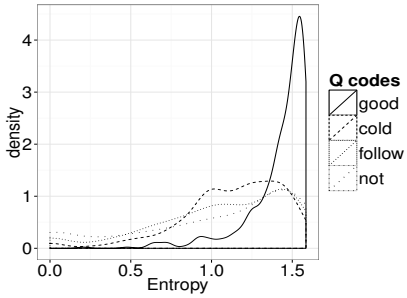

Figure 1: Entropy of Total Duration

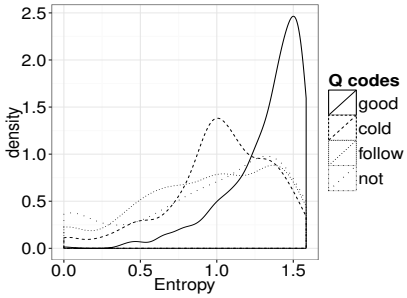

Figure 2: Entropy of Solo Duration

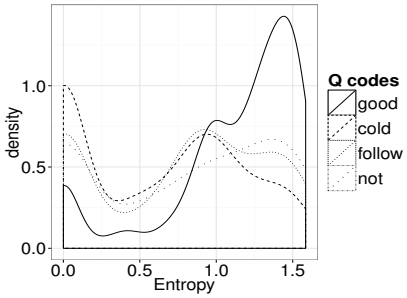

Figure 3: Entropy of Overlap Duration

2.2.2. Prosodic features by individual speaker

We examined a range of features that capture pitch, energy, voicing, and spectral tilt patterns for individual speakers. Unlike the speech activity features, prosodic features in this study were computed "blind" to the prosodic activity of the other participants. Although this approach provides a real-time processing advantage, theoretically we expect performance to be suboptimal since features are not aware of what other speakers are doing. Mapping prosodic features from speaker level to group level is more complex, given that frame-level prosodic information is not binary, as it is for speech activity. In this study, we focus only on the speaker-level prosody features, and leave mapping to the group level for a later study.

Prosodic features were computed for each speaker during each window, and by design did not use any word, phone or other segmentation information. The following features were based on snack get_f0 [16] with default parameter settings: mean and standard deviation of fundamental frequency (f0), standard deviation and maximum of the root mean square energy, and standard deviation and maximum of the peak autocorrelation. In addition, pitch was post-processed [17] to locate f0 maxima and record both pitch peak values and pitch peak locations in time. The latter were used as a crude way to capture the distribution of syllable locations, as an estimate of speaking rate and rhythmicity measures. Features used in classification included the maximum, mean, and variance of the distance between peaks, the distance between the window start and the maximum pitch peak, and the standard deviation of the peak f0 values. Spectral tilt features (n=7) included the mean, standard deviation and maximum values of features described in [18]. The idea was to capture changes in vocal effort, as well as voice quality, that could be used to differentiate dominance, uncertainty or off-talk. For all features, two versions were computed: an un-normalized version and a version normalized by session statistics for the speaker. The latter used mean normalization; percentile normalization gave similar results.

2.2.3. Classifiers

Machine learning experiments were conducted to assess the predictability of collaboration labels. These experiments used both normalized and non-normalized versions of both speech activity features and prosodic features. Group-level features based on speech activity comprised 1623 datapoints with 5 feature dimensions each. Prosodic features were computed independently for each speaker. Since there were 3 speakers per group, there were 3 times as many datapoints (4869 datapoints), as well as a larger set of features (30 dimensions). Predicted values were Q code annotations, annotated at the group level; the codes were replicated at the speaker level for classification experiments involving prosodic features. Since SAD-derived entropy features and many prosodic features depend on the presence of speech (and, in many cases, voicing), there were missing values in both types of feature sets; these were replaced by the mean values of non-missing data.

Classification was performed using the AdaBoost Stagewise Additive Modeling using a Multi-class Exponential loss function (SAMME) algorithm [19]. In total, 10 Random Decision trees were built and fitted to the data by means of boosting [20]. Classification used 5-fold cross validation. For both SAD-derived and prosodic features, folds were created such that no speakers were present in both train and test set partitions.

The AdaBoost algorithm tries to approximate a Bayes classifier by means of an iterative procedure that combines many weak classifiers. Specifically, the AdaBoost algorithm starts by building a classifier with un-weighted training samples. The algorithm then increases the weights for the misclassified data samples and builds a new classifier, taking into account the new weights. This procedure is repeated iteratively. Each classifier is assigned a score, and the final classifier is the linear combination of the classifier from each stage [21]. For multi-class classification, the AdaBoost algorithm can be applied by reducing the multi-class classification problem to multiple two-class problems, or by using the AdaBoost SAMME algorithm [19].

2.3. Results and Discussion

2.3.1. Descriptive statistics on speech activity features

Descriptive results for the speech activity features are shown in Figures 1-3. The figures show Gaussian kernel density estimations of the speech activity entropy distributions, broken down by Q code. We expect that most good collaborations involve roughly equal participation, and thus segments coded as "good" should have high entropy values (close to the maximum of about 1.585). This appears to be borne out in all three plots, in which the distributions of each entropy statistic for "good" collaborations peak near the maximum entropy value. This indicates that during good collaboration windows, students (and all pairwise combinations of students) tend to speak roughly equal amounts. Furthermore, the other collaboration types (including "not") have far lower densities of high entropy values, indicating that when students are speaking equally it is likely, though not certain, that they are collaborating well.

Similarly, for collaborations in which one student is dominating, "follow the leader," we might expect that the "leader" would talk far more than the other two. This expectation would

Table 2: *Class results for F_1 and (Accuracy) by model type. Chance using crude method: assign all tokens to most frequent class.*

	Chance	Model			
Q Code	Brute Force	Prosodic by Speaker (Normalized)	Prosodic by Speaker (Non-Normalized)	Group Speech Activity (Normalized)	Group Speech Activity (Non-Normalized)
Good	48.5% (100%)	38.4% (38.5%)	39.4% (41.1%)	45.7% (52.4%)	52.3% (55.4%)
Cold	0% (0%)	28.8% (29.4%)	28.9% (31.0%)	40.4% (40.6%)	39.7% (41.4%)
Follow	0% (0%)	21.7% (21.0%)	22.6% (22.7%)	22.4% (22.3%)	23.5% (22.6%)
Not	0% (0%)	20.9% (21.8%)	24.7% (24.5%)	27.0% (25.6%)	26.0% (29.0%)
Overall	Chance	Model			
Weighted average	15.7% (32.5%)	29.0% (29.2%)	30.2% (31.3%)	35.8% (37.8%)	37.9% (39.7%)
Unweighted average	12.2% (25.0%)	27.5% (27.7%)	28.9% (29.8%)	33.9% (35.2%)	35.4% (37.1%)

be evidenced by entropy scores near 0 for the Total and Solo Duration statistics for "follow." However, the expectation is not reflected in the data.

For the "out in the cold" class, in which two students are collaborating with each other to the exclusion of the third, we might expect to see this best reflected in the Overlap Duration plot. Specifically, one pair of students (the collaborating pair) should overlap much more than the other two pairs, and we should therefore see an entropy close to 0. This is borne out in Figure 3, where the "cold" distribution peaks at 0.

2.3.2. Classification experiments

Results for classification experiments using AdaBoost are shown in Table 2, using both F_1 and accuracy (shown in parenthesis) as metrics since we are interested in precision and recall irrespective of class priors. The F_1 measure is defined in terms of the harmonic mean of Precision (P) and Recall (R):

$$F_1 = \frac{2PR}{P+R} \tag{2}$$

Results are shown at both the class level (Q codes) and the overall level (weighted and unweighted averages). As a point of comparison to a system without any information, we provide performance for a "brute" force approach that assigns all tokens to the most frequent class–in this case, "good."

As shown, all three models outperform the brute force classifier. The SAD features are better predictors overall than the prosodic features. This result is consistent with the definitions of the four classes, since spoken participation by different numbers of students partly (but not completely) determines annotations. SAD features also take into account time-aligned information about talk patterns of all members of a group, while the current prosodic features know only about one speaker at a time. Despite these limitations, prosodic features still provide information. They give almost the same performance as SAD features for "not" and "follow", and are less biased toward the frequent class than are SAD features. Further gains may be possible by mapping prosodic features to the group level, or if combined with SAD features at the speaker level.

We analyzed feature importance by ranking features according to classifier information gain. Interestingly, the most useful features measured variance and/or maximum values of pitch, variability in spectral tilt, or variation in the location of pitch peaks within the window (correlated with lack of regular pitch accent timing). More analysis is needed, but these features tend to capture a departure from a canonical speaking style. These features may be most important because they help detect off-talk, including stylistic changes associated with joking, strange noises, imitations, teasing, exclamations, and so on. This explanation is consistent with the relatively good performance of prosody for the "not collaborating" class, since off-talk was considered a non-collaborative behavior.

For both feature types, mean-based and percentile-based normalization using the speakers session statistics gave no advantage over the features that were not speaker-session normalized. This result deserves further attention. Since SAD features capture relative amounts of talk over the group, one would expect that normalizing for session would help adjust for differences in idiosyncratic groupings of speakers with different SAD habits. For prosodic features, pitch is typically speaker-normalized, but this was not motivated here. One possible explanation is that the age ranges of the children made this less of an issue than for adult males and females, but more investigation is needed. A second possibility is that better normalization is needed, although speaker-based percentile normalization of pitch should be an appropriate first approximation. A third possibility is that the types of events the features cue could be extreme enough to show up regardless of normalization (for example, features associated with exclamations or off-talk).

3. Conclusions

Collaborative learning is an important skill for students to master, and there is a clear need to assist teachers in monitoring simultaneous group learning in the classroom. Our findings offer first results which suggest that automatic processing of speech offers promise for addressing this important need. We used simple features based only on speech activity and speech prosody. This approach is privacy-preserving, since it uses no video or lexical information.

Results reveal that both SAD and prosodic features perform well above chance, and offer potential complementarity based on class-specific results. Furthermore, session-based normalization at the speaker level did not appear to be necessary, suggesting good performance may be achieved without look-ahead. Future work will address the mapping of prosodic features to the group level, and assessment of whether further gains are possible using group-level fusion of the two feature types.

4. Acknowledgements

We gratefully acknowledge the contributions and support of Diana Jang, Erik Kellner, Tiffany Leones, Tina Stanford, Jeremy Fritts and Jeremy Roschelle. This work was funded by National Science Foundation grant #DRL-1432606.

5. References

[1] National Research Council, *Assessing 21st century skills: Summary of a workshop*, Washington, D.C.: The National Academies Press, 2011.

[2] G. W. Ladd, B. Kochenderfer-Ladd, K. J. Visconti, I. Ettekal, C. M. Sechler and K. I. Cortes, "Grade-School Childrens Social Collaborative Skills: Links With Partner Preference and Achievement," *Am. Educ. Res. J.*, vol. 51, no. 1, pp. 152-183, 2013.

[3] E. G. Cohen, "Restructuring the Classroom: Conditions for Productive Small Groups," *Rev. of Educ. Res.*, vol. 64, no. 1, pp. 1-35, 1994.

[4] G. Erkens and J. Janssen, "Automatic coding of dialogue acts in collaboration protocols," *Int. J. Comput. Collab. Learn.*, vol. 3, no. 4, pp. 447-470, 2008.

[5] B. M. McLaren, O. Scheuer, M. De Laat, R. Hever, R. De Groot and C. P. Rosé, "Using machine learning techniques to analyze and support mediation of student e-discussions," *Frontiers in Artificial Intelligence and Applications*, vol. 158, no. 1, pp. 331-338, 2007.

[6] C. P. Rosé, Y. C. Wang, Y. Cui, J. Arguello, K. Stegmann, A. Weinberger and F. Fischer, "Analyzing collaborative learning processes automatically: Exploiting the advances of computational linguistics in computer-supported collaborative learning," *Int. J. Comput. Collab. Learn.*, vol. 3, no. 3, pp. 237-271, 2008.

[7] M. Worsley and P. Blikstein, "Towards the development of multimodal action based assessment," *LAK 13 Proc. Third Int. Conf. Learn. Anal. Knowl.*, pp. 94-101, 2013.

[8] G. A. Gweon, P. B. Agrawal, M. C. Udani, B. A. Raj and C. P. Rose, "The automatic assessment of knowledge integration processes in project teams," in *Connecting Computer-Supported Collaborative Learning to Policy and Practice: CSCL 2011 Conference Proceedings - Long Papers*, vol. 1, pp. 462-469, 2011.

[9] G. Gweon, M. Jain, J. McDonough, B. Raj and C. P. Ros, "Measuring prevalence of other-oriented transactive contributions using an automated measure of speech style accommodation," *Int. J. Comput. Collab. Learn.*, vol. 8, pp. 245-265, 2013.

[10] D. Kuhn, "Thinking together and alone," *Educ. Res.*, vol. 44, no. 1, pp. 46-53, 2015.

[11] A. Tsiartas, T. Chaspari, N. Katsamanis, P. K. Ghosh, M. Li, M. Van Segbroeck, A. Potamianos and S. Narayanan, "Multi-band long-term signal variability features for robust voice activity detection," in *Interspeech 2013*, pp. 718-722, 2013.

[12] Ghosh, P. Kumar, A. Tsiartas and S. Narayanan, "Robust voice activity detection using long-term signal variability," *Audio, Speech, and Lang. Proc., IEEE Trans.*, vol. 19, no. 3, pp. 600-613, 2011.

[13] H. Hung, Y. Huang, G. Friedland and D. Gatica-Perez, "Estimating Dominance in Multi-Party Meetings Using Speaker Diarization," *IEEE Trans. on Audio, Speech, and Lang. Proc.*, vol. 19, no. 4, pp. 847-860, 2011.

[14] R. Rienks and D. Heylen, "Dominance Detection in Meetings Using Easily Obtainable Features," *Machine Learning for Multimodal Interaction*, vol. 3869, pp. 76-86, 2006.

[15] C. Shannon, "A Mathematical Theory of Communication," *Bell System Technical Journal*, vol. 27, no. 3, pp. 379-423, 1948.

[16] K. Sjölander-SNA, "The Snack Sound Toolkit," *Speech.kth.se*, 2015. [Online]. Available: http://www.speech.kth.se/snack/.

[17] N. C. Yoder, "PeakFinder (Matlab program)," *Mathworks.com*, 2015. [Online]. Available: http://www.mathworks.com/matlabcentral/fileexchange/25500-peakfinder-x0–sel–thresh–extrema–includeendpoints–interpolate-.

[18] E. Shriberg, A. Stolcke and S. Ravuri, "Addressee Detection for Dialog Systems Using Temporal and Spectral Dimensions of Speaking Style," *Proc. Interspeech*, 2013.

[19] J. Zhu, H. Zou, S. Rosset and T. Hastie, "Multi-class adaboost," *Statistics and Its Interface 2*, vol. 3, pp. 349-360, 2009.

[20] T. Hastie, R. Tibshirani and J. H. Friedman, *The elements of statistical learning*, New York: Springer Verlag.

[21] Y. Freund and R. Schapire, "A decision theoretic generalization of on-line learning and an application to boosting," *Jour. of Comp. and Sys. Sci.*, vol. 55, pp. 119-139, 1997.