Mapping Individual to Group Level Collaboration Indicators Using Speech Data

Cynthia M. D'Angelo, University of Illinois at Urbana-Champaign, cdangelo@illinois.edu
Jennifer Smith, SRI International, jennifer.smith@sri.com
Nonye Alozie, SRI International, maggie.alozie@sri.com
Andreas Tsiartas, SRI International, andreas.tsiartas@sri.com
Colleen Richey, SRI International, colleen.richey@sri.com
Harry Bratt, SRI International, harry.bratt@sri.com

Abstract: Automatic detection of collaboration quality from the students' speech could support teachers in monitoring group dynamics, diagnosing issues, and developing pedagogical intervention plans. To address the challenge of mapping characteristics of individuals' speech to information about the group, we coded behavioral and learning-related indicators of collaboration at the individual level. In this work, we investigate the feasibility of predicting the quality of collaboration among a group of students working together to solve a math problem from human-labelled collaboration indicators. We use a corpus of 6th, 7th, and 8th grade students working in groups of three to solve math problems collaboratively. Researchers labelled both the group-level collaboration quality during each problem and the student-level collaboration indicators. Results using random forests reveal that the individual indicators of collaboration aid in the prediction of group collaboration quality.

Introduction and theoretical background

Management and assessment of collaborative learning tasks is difficult in typical classrooms when teachers attempt to monitor 10-15 groups with 2-3 students in each group. Teachers need tools to help them identify groups of collaborating students that are doing well and those that need help (and what kind of targeted help they need). Ideally, teachers would listen to peer interactions in each group for long enough to understand how discourse is proceeding. However, teachers cannot listen to more than one group at a time and therefore this in-depth listening, evaluation, and feedback they provide cannot be done at scale, even within a single classroom with more than a few groups (Kaendler et al., 2014). Students talking to one another in face-to-face environments is the natural setting for classroom collaboration, and so speech-based solutions are needed to address this problem. We see speech-based analytics functioning to not replace the teacher's role, but rather to inform the teacher's exploration of group dynamics, diagnosis of issues, and development of intervention plans.

In this work, we investigate the feasibility of using students' speech during collaborative activity to determine group collaboration quality. While speech activity alone is indicative of some basic features of good/productive (or bad/unproductive) collaboration (e.g., one person is doing all of the speaking in a group or one person never talks), some key information is left out that is relevant to assessing collaboration quality. What students are talking about, and the function of that speech, plays a large role in determining collaboration quality. One student could have been talking about their plans for the weekend or just automatically agreeing with everything the other two people said without adding anything new. Productive collaboration requires the participants to directly engage in one another's thinking, which includes listening and responding to the others in the group (Kuhn, 2015). Automatically detecting this type of complex thinking and behavior will likely require more than just speech activity information. Incorporating individual indicators of collaborative actions, based on what individuals are doing to contribute to the group's problem solving, should help assess the overall group collaboration quality. One major challenge is how to effectively map individual contributions along a quickly moving timescale to the overall collaboration of the group at specific time points.

Methods

Data collection

Data collection took place in the spring of 2015. 134 middle school students (67 in sixth grade, 35 in seventh grade, and 32 in eighth grade; 68 female, 66 male) from six different schools participated. Students worked on the collaborative activities after school. 80 collaborative sessions were recorded (about 15-20 minutes each). Most students participated in two sessions with different group configurations. Each student had an individual head-mounted microphone and the whole group was video recorded.

Participating students worked together in groups of three on a set of collaborative math activities on iPad stations. The collaborative math activities included 12 items, where each item required the three students to work together and talk to each other to coordinate their three answers to the problem. The tasks are similar to those described by Roschelle and colleagues (2010). These tasks incorporate two important principles for effective collaborative learning tasks: positive interdependence (students must each contribute to the overall task for it to succeed) and individual accountability (students cannot succeed by freeloading) (Roschelle et al., 2010).

Collaboration coding schemes

All audio recordings made during active work on the mathematics problems were manually annotated by education researchers for (1) indicators of collaboration performed by individual students and (2) the overall collaboration quality of the group for each item. Annotators assigned collaboration indicator codes (or "I-codes"), each with a start and end time, to the individual audio channels. The I-codes are: Monitoring the progress of the group, Verbalizing their thoughts, Reading the problem aloud, Communicating that they are thinking, Making plans to solve a problem, Turn sharing, Ignored speech, Acknowledging another student, Explaining the problem to the group, Expressing lack of understanding, Giving away an answer, Inviting others to contribute, Asking a question, Agreeing with another student, and Disagreeing with another student.

A separate group of researchers annotated the data set for group-level collaboration quality codes (or "Q-codes"). These codes represented the degree to which the three students collectively were engaging in good collaboration. Importantly, they depend not on how much student talked but on whether and how much each student was engaged intellectually in the group problem solving. Annotators assigned collaboration quality codes at the item level which varied in length depending on how long it took to solve the problem.

The Q codes are: *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 others), *not collaborating* (no students are actively contributing to solving the problem (either off-task or independently working). More information about the coding procedures, the coding schemes, and the descriptive statistics of the Q-codes and I-codes can be found in Richey and colleagues (2016).

Features

Making predictions about the collaboration quality of a group of students by using individual student speech requires some form of mapping from the information contained in the individual speech signals to information about the group. In previous papers (Basiou et al., 2016; Smith et al., 2016), we explored the mapping of features derived from speech activity detection and acoustic, prosodic, and temporal information extracted from the speech signal. We found that features extracted from speech activity detection (SAD) produced the best classification results. In this paper, we explore a different approach: first coding the speech with behavioral and learning-related indicators of collaboration (I-codes) at the individual level and mapping these I-codes to the group-level Q-codes in different ways.

Features derived from collaboration indicator codes

We extracted several features to aggregate the I-codes assigned to the individual students' speech to the group level for each item that the group solved. The first set of features we computed capture information about the frequency of the codes, the co-occurrence of the codes, and the number of students per group with each code.

We hypothesized that the frequency of certain codes would correlate with collaboration quality. Due to sparsity among many of the I-codes, using frequency codes for each student was unfeasible. Thus, looked at group-level frequency along with a count of the number of students with that code. We also hypothesized that the co-occurrence of the I-codes within an item might provide important information. For example, if both *verbalize* and *agree* occur within a item, then we know that at least one student expressed their thought process to the group and at least one student expressed agreement. However, due to I-code sparsity, we found that the combination of the co-occurrence features, the frequency of each I-code, and the count of students who received each code was not enough information for our model to predict collaboration quality above chance.

Upon further inspection of the I-codes, we realized that we needed to account for variation in both the length of the items and the length of the I-codes. Some items were very short (less than a minute) and others were quite long (six or more minutes). Thus, we created a length-normalized version (i.e., codes per minute within the item) of the counts of each I-code.

We also hypothesized that I-code duration may be important for prediction. For example, a student who *verbalizes their thinking* for one second might say: "I think the speed should be five", while a longer *verbalize* utterance could be: "Since speed is distance divided by time, shouldn't we be dividing 15 by 3? That would

make 5." The latter utterance should be weighted differently from the former. We computed the relative duration of each I-code by summing the total duration of each I-code per channel and dividing by the length of the item. We also ordered these item-length-normalized-durations by the students' talkativeness. This resulted in codes like "the proportion of the item that the least talkative student spent verbalizing."

Speech activity features

The data collection setup allowed students to speak freely, resulting in audio recordings with overlapping speech from the three students. To address this issue, we used a speech activity detection (SAD) system based on speech variability (Ghosh, Tsiartas, & Narayanan, 2013) and ran the system independently on each of the three student channels. Using the segmental and durational information resulting from the speech activity segmentation, we extracted SAD-based features that characterized either the individual student speech or the speech patterns of the group. The features capture information about the number, duration and location of the speech regions, similar to those used in studies of dominance in multi-party meetings (e.g., Hung et al., 2011).

Specifically, we extracted features that capture information about the overall speech duration and the "spurts" of speech. Spurts were defined as regions of speech that are at least 50ms long and were uninterrupted by pauses longer than 200ms. As students deal with the cognitive load of simultaneously solving math problems and negotiating with their peers, they frequently interrupt each other or speak in short incomplete phrases.

We extracted these SAD-based features across the channels individually and in combination, taking into consideration speech activity from regions in which: each individual student was the only speaker, each individual student spoke, ignoring speaker overlap, each pair of students spoke simultaneously, all students were silent, or all students spoke simultaneously.

Duration-based features for individual and pairs of students were mapped to the group level using ratios and entropy of the statistics as described in Smith and colleagues (2016). These features capture information about the distribution of speech duration across the members of the group. The spurt-based features capture information about the turn-taking and other features of the speaking style of the group.

Machine learning experimental setup

Feature sets

Classification experiments were run on multiple different feature sets. The first was "I-Codes-1": the initial set of I-code-derived features comprised of the counts of each I-code, co-occurrences of each pair of I-codes, and the number of students with each I-code. Next, we ran experiments on "I-Codes-2": the set of I-codes that took the length of the item and I-code duration into account. Specifically, the item-length-normalized counts of each I-code, the relative durations of each I-code (ordered by talkativeness), and the ratios of the relative durations. We also tried a feature-level-fusion of the two I-code sets, which we call "I-Codes-1+2". For comparison, we ran classification using the SAD feature set, comprised of all the duration and spurt based features derived from speech activity detection. We also fused the SAD and I-Codes-1+2 sets at the feature level.

Classifier

We used random forest models (RF) (Breiman, 2001) to classify the quality of collaboration from our features. The individual decision trees are created by selecting a random sample with replacement of the training set, along with a random sample of the variables and fitting a tree to that subset. This process makes for a collection of weak classifiers that are less susceptible to over-fitting than a single decision tree. In this study, we tuned the number of variables ("n_var") used to build each tree using the training split of each fold.

Classification was performed using leave-one-day-out cross validation to ensure that no student(s) were in both the training and the testing sets. This cross validation set up was chosen because, for each day of the data collection, most students were in two different groups. Thus, the only way to ensure that there was no speaker overlap between the training and testing set was to leave out an entire day. This resulted in 21 folds.

We built multiple RF models for each of the training sets in the 21 folds, varying the n_var parameter and then repeating 10-fold cross-validation three times for each value of n_var. The accuracy from each model within a fold was used to select the best model for that training set. Then, the corresponding test set was used to calculate classification weighted and unweighted accuracy and F1-scores.

Findings

Collaboration quality prediction was investigated using I-Codes-1, I-Codes-2, I-Codes-1+2 and SAD features. A baseline system, calculated by always predicting the most frequent class, was the comparison. We also used the SAD features as a baseline, because they were the best performing features in (Smith et al., 2016).

All feature sets derived from collaboration indicators outperform chance and the I-Codes-2 set also outperforms the SAD derived features in UWA. UWA takes into account the accuracy of prediction across classes and is a better measure of performance for this application because teachers will be concerned with accurately predicting all types of collaboration, not just the dominant class, "good collaboration".

In terms of UWA, the I-Codes-1 and I-Codes-2 set achieve a 36.0% and 72.8% relative improvement over Baseline, respectively. The I-Codes-2 set also outperforms the SAD derived features, with a 11.9% relative improvement. The main difference between the two sets is that the former does not take I-code duration or item duration into account. I-code duration information may be especially important in this problem due to variability in the style of the human annotators; the start and end times of the coded utterances were hand selected by the annotators and the pauses in speech may have been treated inconsistently. The I-Codes-2 set also featured item-length-normalization which is crucial due to the variable length of time it took students to complete the items.

Fusing I-Codes-1 and I-Codes-2 ("I-Codes-1+2") did not produce an increase in accuracy. This is potentially due to increasing the dimensionality of the feature set. Moreover, using both I-Codes-1 and I-Codes-2 feature sets may not offer additional information. We also calculated the results of collaboration quality prediction using feature level fusion of the best predictors. The results indicate that fusing the I-codes with the SAD features give a ~5% relative improvement in terms of weighted accuracy, but no improvement in UWA.

Conclusions and implications

Classification results align with the education literature and indicate that the I-codes are an important speech-based indicator of collaboration. Thus far, we have explored modeling of the I-codes that takes item length and duration into account, but ignores additional dialogue-level information, such as turns, speaker sequence, and I-code sequence. Future work will explore new ways of modeling I-codes to further improve the classification performance.

We plan to evaluate additional transformations and methods to convert the I-codes to meaningful information for collaboration prediction. We also plan to evaluate the effect of reducing the dimensionality of the codes by clustering the I-codes. In addition, we plan to extract new information from the dialogue between the students, such as turn-taking and sequence modeling of the I-codes. Moreover, we plan to explore additional speech features that can describe additional verbal and non-verbal speech behaviors.

References

Basiou, N., Tsiartas, A., Smith, J., Bratt, H., Richey, C., Shriberg, E., D'Angelo, C., & Alozie, N. (2016). Privacy-preserving speech analytics for automatic assessment of student collaboration. *Proceedings of the Speech Prosody Conference 2016*. Boston, MA.

Breiman, L. (2001). Random forests. *Machine learning, 45*(1), 5–32.

Ghosh, P. K., Tsiartas, A., & Narayanan, S. (2011). Robust voice activity detection using long-term signal variability. *IEEE Transactions on Audio, Speech, and Language Processing*, *19*(3), 600–613.

Hung, H., Huang, Y., Friedland, G., & Gatica-Perez, D. (2011). Estimating dominance in multi-party meetings using speaker diarization. *IEEE Transactions on Audio, Speech, and Language Processing*, *19*(4), 847-860.

Kaendler, C., Wiedmann, M., Rummel, N., & Spada, H. (2014). Teacher Competencies for the Implementation of Collaborative Learning in the Classroom: a Framework and Research Review. *Educational Psychology Review*, *27,* 505-536.

Kuhn, D. (2015). Thinking together and alone. *Educational Researcher*, *44*(1), 46-53.

Richey, C., D'Angelo, C., Alozie, N., Bratt, H., & Shriberg E. (2016). The SRI Speech-Based Collaborative Learning Corpus. *Proceedings of the Speech Prosody Conference 2016*. Boston, MA.

Roschelle, J., Rafanan, K., Bhanot, R., Estrella, G., Penuel, B., Nussbaum, M., & Claro, S. (2010). Scaffolding group explanation and feedback with handheld technology: Impact on students' mathematics learning. *Education Tech Research Dev, 58*, 399-419.

Smith, J., Bratt, H., Richey, C., Bassiou, N., Shriberg, E., Tsiartas, A., D'Angelo, C., & Alozie, N. (2016). Spoken interaction modeling for automatic assessment of collaborative learning. *Proceedings of the Speech Prosody Conference 2016*, pp 277-281. Boston, MA.

Acknowledgments

This material is based upon work supported by the National Science Foundation under Grant No. DRL-1432606. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.