GRASP: The Group Learning Assessment Platform

Gahgene Gweon, Rohit Kumar & Carolyn Penstein Rosé,
Carnegie Mellon University, 5000 Forbes Avenue, Pittsburgh, PA 15213
Email: {ggweon,rohitk,cprose}@cs.cmu.edu

Abstract: We demonstrate a prototype assessment technology designed to enable unobtrusive, real time assessment of group dynamics from speech. As part of that work, we describe a test bed for experimentation with alternative approaches for automatic processing of speech for this purpose. Furthermore, we present a specific successful technique for predicting activity levels and amount of overlapping speech in recordings of actual student group meetings recorded over a semester of a graduate engineering design project course.

Introduction

Multi-disciplinary design project classes present challenges both for supporting and for assessing learning because the learning is self-directed and knowledge is acquired as needed throughout the design process. What makes it especially tricky from an instructor perspective is that the bulk of student learning takes place without the instructor present. While this provides students with opportunities to develop skills related to "learning to learn", it can also mean that instructors are left not knowing when and how they can intervene to support the students most effectively. It is well known from the social psychology literature on group work that groups frequently do not function in an ideal way (e.g., Faidley et al., 2000).

Recently, in order to address this problem, there have been a number of efforts to support instructors in managing group work by offering them forms of automatic assessment and reporting (e.g., Soller & Lesgold, 2003; Kay et al., 2006; Pianesi et al., 2008). In prior work, researchers have looked at automatically detecting various aspects of student activities during group work (e.g., Kay et al., 2006; Pianesi et al., 2008). Various forms of data have been used including message board postings (Kim et al., 2007), chat data (Soller & Lesgold, 2003), video (Chen, 2003), and audio (DiMicco et al., 2004). Because our goal is to support students in project based courses, speech recorded using mini digital recorders that are small enough for students to carry with them is the most natural medium for our work since student group working meetings are typically conducted face-to-face, sometimes planned and sometimes spontaneously, in any of a number of locations.

We present the Group Learning Assessment Platform (GRASP)[1], a technology designed to enable unobtrusive, real time assessment of group dynamics from digital recordings of student speech. We will begin by describing GRASP at a conceptual level. We then present techniques for predicting activity levels and amount of overlapping speech in recordings of actual student group meetings recorded over a semester of a graduate engineering design project course. We conclude with findings related to our automatic assessment technology as well as our current directions.

Overview of GRASP

The problem of automatic assessment of group learning processes is large and multifaceted. First, different instructors may have a plethora of alternative assessment goals ranging from goals for learning, to goals for social development, to goals for productivity and project success (Gweon, 2008). Assessment criteria must then be operationalized in a way that is reliable and valid. Within the recorded conversational data are a succession of different types of events, which may be catalogued in a variety of ways. As part of the process of doing the automatic assessment from the speech signal, these events must be detected and used to compute indicators related to the selected assessment categories. Finally, the recorded speech itself is multifaceted, encoding both content features related to what was said as well as style and intonation features that indicate how the speech was uttered. All of these types of speech features may be informative, and none of them are trivial to detect. Because of the wide space of possible assessment frameworks that fit this paradigm of detecting events in speech in order to compute indicators that correlate with desired assessment categories, we developed GRASP to be both as a prototype assessment technology and a test bed for exploring alternative assessment approaches.

Figure 1 illustrates the GRASP framework. In stage 1, students carry headphones and mini digital recorders with them so that whenever their project teams meet in any location, they can record their meetings, with the speech from each student recorded in a separate file. In the next stage, we preprocess the speech to create a representation consisting of feature-value pairs we can extract from the speech files. Next we compute indicators from the speech recordings, both related to an individual's participation in a meeting and a group's well functioning, based on comparisons of the speech for the individuals present in the meetings. Using regression models trained over speech data paired with human assessment ratings, we are able to make automatic predictions about how humans would rate group meetings along selected assessment dimensions.

[1] This research was supported in part by NSF Grant EEC-064848.

Finally, these automatic predictions can be displayed for an instructor to observe. This process model was designed to be general, allowing us to explore all four important dimensions in our ongoing work.

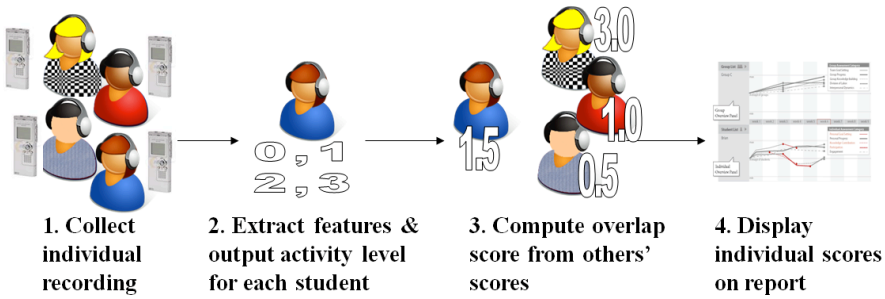

Figure 1 Overview of the four stage automatic assessment process.

Processing the Speech

Using current speech processing technology, one can make use of features that can be extracted from speech, such as pitch and energy level, that say something about the nature of the interaction. In prior work, Dabbs and Ruback shows the usefulness of style of speech in gaining insight into group processes that occur among individuals who participated in group work (Dabbs & Ruback, 1987). In our work, using recordings collected from students during project group meetings, we computed amount of activity level for each student using machine learning technology. From the predicted activity level, two types of measurements were computed: namely, average of the percentage of time when that student was talking during group meetings (average activity level) and average percentage of group mates who were talking during the time when that student was talking (amount of overlap). Average activity level is an approximate measure of the amount of talk that the student contributed during in group meetings. Amount of overlap says something about how seriously a student's group mates take his/her contributions. If overlap is high, then it may be the case that the student's group mates don't find it valuable to stop and listen when he/she speaks.

Before amount of talk and amount of overlap can be computed from speech, it must be segmented, and each segmented must be coded for the amount of speech by the associated student that was detected in it. We chose to segment the speech into 10 second intervals so that it would be reasonable to assume that for most segments, there would be at most a single dominant speaker. We adopted the following 4-point scale for activity level: 0 - no speech from primary speaker; 1 - primary speaker only does back-channeling, where back-channeling is a way of showing a speaker that you follow and understand their contributions, often through interjections such as , "I see", "yes", "OK", "uh-huh"; 2 - primary speaker speaks holds the floor for less than half of the 10 seconds; 3 - primary speaker speaks holds the floor for more than half of the 10 seconds

We first verified that human annotators could make this judgment reliably from the audio recordings of individual segments. Using this coding scheme, the inter rater reliability evaluated for two coders over 144 segments was 0.78 Kappa. With the reliable coding scheme, a single coder then coded 1132 segments (distributed evenly across students from a project course). The largest proportion of segments was coded as 0, which amounted to 47.5% of the segments. 8.5% were coded as 1, 30.5% as 2, and 13.5% as 3.

In order to apply machine learning to speech, each segment of speech must first be transformed into a set of feature-value pairs. A total of 39 features were extracted for each of the 10 second segments using wavesurfer (Beskow & Sjlander, 2000). These features are comprised of structural aspects of speech, features related to F0 and power. With the coded speech data after it had been transformed into a vector representation, we then evaluated whether it was possible to use machine learning to automatically assign segments of speech to one of these four categories with high enough accuracy. We used Weka's SMO learning algorithm (Witten & Frank, 2005). In order to avoid the evaluation results being inflated due to overlap in speakers between train and test sets, we adopted a cross-validation evaluation methodology where a model was first trained on all but one student, and then performance was evaluated over the segments of the remaining student. We did this for each student and then averaged across students to compute the performance of 74.26% accuracy. We then validated the model by using the human coded numbers for each student to compute an average activity level, and then made a similar computation using predicted values from the cross-validation experiment. When we correlated the average activity levels for each student based on human codes with those based on the automatic codes, we achieved a correlation coefficient of 0.97, indicating that we can achieve a reliable estimate of activity level using a machine learning model. We then trained a model using all of the coded data, which we used in the subsequent analysis we discuss in this poster. We applied the trained model to a separate set of speech data from that used to build the models for predicting student activity. Altogether 18 students' recordings were segmented into 10 second segments. The length of each recording differed due to differences in meeting lengths. The number of segments ranged between 7 minutes 30 seconds to 2 hours 19 minutes 50 seconds in

length (45 to 839 ten second segments), with an average of 47 minutes in length (282 segments). Using the speech model just described, student recordings were assigned amount of talk values. Example predictions are shown in table 1, where we see that Student 1 is the dominant speaker for all three segments shown. Students 2 and 3 start to contribute more substantially during the third segment, and student 4 only does back-channeling.

Table 1: Example predictions of speech activity. Numbers in parenthesis indicate the smoothed values.

(sec)	Student1	Student2	Student3	Student4
0~9	3 (3)	0 (0)	0 (0.5)	1 (0.5)
10~19	3 (3)	0 (2/3)	1 (1)	0 (2/3)
20~29	3 (3)	2 (1)	2 (1)	1 (1/3)
30~39	3 (3)	1 (1.5)	0 (1)	0 (0.5)

Using the same predictions of activity level per segment, an amount of overlap index was calculated for each student. Overlap is defined as the amount of activity level by group mates when the student is actively talking. In order to compensate for some error in coding activity level, we first smoothed the predictions of activity level by averaging the activity level prediction of a segment with those of the segment before and the segment after. The resulting smoothed scores were then real values between 0 and 3. We then applied a threshold to determine which segments we would treat as segments during which a student was speaking. The threshold for each meeting was computed as the average of all the activity level over all the smoothed segments in that meeting. For each of the 10 second segment, we compared the student's smoothed activity level to the threshold. Therefore, if a student's smoothed activity level in the given segment was above the threshold, that student was considered as speaking during that segment. Next, for the segments where activity level was larger than the threshold, which are considered as segments where the student was talking, we computed the average smoothed activity level of the other meeting participants during that segment. We consider this the overlap score for that student, which indicates the prevalence of other group members talking at the same time when this student is talking. Finally, after computing amount of overlap for all the 10 second segments, an average of the amount of overlap is computed over all the segments in a given meeting to yield one overlap score per student.

Current Work

We have described a technique for predicting amount of overlapping speech in recordings of group meetings. Results from regression models trained to predict assessment categories from our prior work (Gweon, 2008) with activity level and overlap indicators suggest that these indicators are useful for identifying the students that instructors are more likely to make faulty assessments of based on in class observations. Our hope is that this insight might overcome "blind spots" that instructors might have in order to improve the support they can offer students. One important next will be to test how providing such information influences instructors' behavior, well functioning of student groups, and ultimately student learning and project success. Further work can also be done to identify other quantities that can be extracted from the speech that might be useful for project course instructors. Our research group is planning additional data collection efforts in another types of project courses in order to enable the development of a more generally useful framework for assessment of student groups.

References

Dabbs, J., and Ruback, B. (1987). Dimensions of group process: Amount and structure of vocal interaction. *Advances in Experimental Social Psychology, 20*, pp 123-169.

DiMicco, J., Pandolfo, A., and Bender, W. (2004). Influencing group participation with a shared display. *Proc. CSCW 2004*, ACM Press, 614-623.

Faidley, J., et. al. (2000). How are we doing? Methods of assessing group processing in a problem-based learning context. In Evensen, D. H., and Hmelo, C. E. (eds.), *Problem-Based Learning: A Research Perspective on Learning Interactions*, Erlbaum, NJ, 109-135.

Gweon, G. (2008). Predicting Group Behavior from Audio Recordings of Meetings. In *Proceedings of ACM-SIGCHI Doctoral Consortium.*

Kay, J., Maisonneuve, N., Yacef, K. & Reimann, P. (2006). *Wattle Tree: What'll It Tell Us?*, University of Sydney Technical Report 582, January 2006.

Kim, J., Shaw, E., Chern, G, and Herbert, R. (2007). Novel tools for assessing student discussions: Modeling threads and participant roles using speech act and course topic analysis. *In proc. AIED*, 2007.

Pianesi, F., et. al. (2008). Multimodal support to group dynamics. *Personal and Ubiquitous Computing.* Vol 12, No 3., 181-195.

Soller, A., Lesgold, A. (2003). A computational approach to analyzing online knowledge sharing interaction. In Proc. AIED 2003, 253-260.

Witten, I. H., Frank, E. (2005). *Data Mining: Practical machine learning tools and techniques*, Morgan Kaufmann