The Automatic Assessment of Knowledge Integration Processes in Project Teams

Gahgene Gweon, Carnegie Mellon University, 5000 Forbes Ave, Pittsburgh, PA, ggweon@cs.cmu.edu
Pulkit Agrawal, Indian Institute of Technology, Kanpur, Uttar Pradesh 208016, India, pulkit@iitk.ac.in
Mikesh Udani, Indian Institute of Technology, Kharagpur, West Bengal 721302, India, mikesh.iitkharagpur@gmail.com
Bhiksha Raj, Carnegie Mellon University, 5000 Forbes Ave, Pittsburgh, PA, bhiksha@cs.cmu.edu
Carolyn Rose, Carnegie Mellon University, 5000 Forbes Ave, Pittsburgh, PA, cprose@cs.cmu.edu

Abstract: Automatic assessment of group processes in collaborative groups is one of the holy grails of the computer supported collaborative learning community. The conversation in collaborative work provides an important window into the inner workings of a group. In this paper we present work towards detecting where students are displaying "reasoning" in conversational speech and how others are building upon those expressions of reasoning ("idea co-construction (ICC)"). Such technology would add to the body of work in educational data mining another means of monitoring student work as well as contributing to the area of automatic collaborative process analysis. We begin by discussing our operationalization of targeted group processes, namely reasoning and ICC. We then discuss the level of success we are able to achieve applying machine learning technology to replicate this human analysis using simple audio signal processing techniques.

Introduction

As communication technologies such as cell phones and voice over IP become more ubiquitous and allow for communication and collaboration over multiple modalities including video, audio, and text to be accessible any time and any place, the line between online group learning and face-to-face group learning begins to blur. Furthermore, as more and more collaboration takes place over video and audio channels, the need grows for the CSCL community to think about how to extend collaboration support technologies from the text realm into audio and eventually video. In this paper we present work towards assessment of group processes from speech data, specifically focusing on processes related to knowledge transfer and knowledge integration within groups. Specifically, we target design project classes, which 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 regardless of whether instructor supervision takes place online, in a whole class setting, or in face-to-face advising meetings with student groups, 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).

Prior work investigating assessment practices of project course instructors reveals both the importance and difficulty of accurately assessing important group processes (Gweon et al., 2011). In this work, project course instructors reported attempting to assess groups in terms of planning and goal setting, productivity and progress, knowledge sharing and group knowledge integration, leadership and division of labor, and interpersonal dynamics. Because each student's expertise and experience provides an important added dimension to the project development process in a multi-disciplinary team, at the level of knowledge sharing and group knowledge integration, instructors wanted to see students behaving as intellectual leaders within their groups, taking the initiative to contribute their own unique expertise and perspective to the group. Beyond that, they wanted to see the contributed ideas taken up and transformed by the group as evidence that the end product would represent a true integration of expertise across the students within the group, and not just a patchwork product that frequently results from dysfunctional group efforts.

In our work we operationalize group knowledge integration processes through an adaption of the construct of transactivity from the field of collaborative learning (Berkowitz & Gibbs, 1983; Teasley, 1997; Weinberger & Fischer, 2006), which we refer to as Idea Co-Construction (ICC). Transactivity has its foundations in Piaget's theory of learning, and is theorized to provide opportunities for cognitive conflict to be triggered within group interactions, which may eventually result in cognitive restructuring (de Lisi & Golbeck, 1999). The construct of transactivity makes several important distinctions related to the notion of intellectual leadership and group knowledge integration that are relevant for our work. First, is the important distinction between contributions that visibly display reasoning behind an assertion versus contributions where the reasoning that lead to an assertion is hidden. Assertions that display reasoning are further subdivided into ones

that build on or operate on prior assertions, which are the transactive variety, and ones that represent a new direction in the conversation, which are externalizations. Externalizations position students as intellectual leaders within a conversation. However, true leadership requires that the leader is received as such by the other group members. Thus, externalizations that are not followed by transactive contributions building on them may be regarded as failed attempts at intellectual leadership. A more complete picture of intellectual leadership within a group can be obtained through tracking the distinction between assertions that do not make reasoning visible, externalizations, and idea co-construction. Our notion of ICC is different from earlier characterizations of transactivity. For example, our notion of ICC does not necessarily require that expressions of reasoning involve a comment on or operation on articulated reasoning that was previously contributed. Instead, the expression of reasoning may simply integrate information articulated previously even information articulated in such a way that it does not conform to our operationalization of reasoning. We adopt this slightly more relaxed notion of "building" in order to be more inclusive of the types of integrative contributions that students contribute since in our experience, the ideal of transactivity is not frequently achieved.

In our work we build on recent efforts to support instructors in managing groups by offering them forms of automatic assessment and reporting. In prior work, researchers have looked at automatically detecting various aspects of student activities during their work together (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). Our contribution is technology to use speech data to distinguish between ICC, externalizations, and contributions that do not display reasoning. Thus, in the remainder of the paper we first situate our work in the midst of current directions in speech processing. Next we discuss our approach to operationalizing reasoning displays and ICC. We then move on to a discussion of the technological contribution of the paper. Finally, we present an evaluation of our approach and conclude with a discussion of current directions.

Motivation and Background

Automatic analysis of ICC and related constructs such as transactivity is not a completely new direction in the CSCL community in itself. However all of the prior published work was related to automatic processing of text, such as newsgroup style interactions (Rosé et al., 2008), chat data (Joshi & Rosé, 2007), and transcripts of whole group discussions (Ai, Kumar, Nguyen, Nagasunder, & Rosé, 2010). One lesson learned by comparing across efforts to detect transactivity in a variety of types of interactions is that a key feature enabling high accuracy of recognition is being able to measure content similarity between a contribution and the contributions from other conversational participants that occurred within the same topic segment earlier within the conversation. For example, Rosé and colleagues (2008) report that in a classification task with a coding scheme related to transactivity, adding a single feature representing content similarity with prior contributions within the same thread from other participants to a baseline feature space, keeping all other aspects of the modeling technique constant, produced an increase in agreement with human coding from 0.5 Kappa to 0.69 Kappa.

The unique contribution of the work presented here is that it is not applied to text, but to recorded speech. Although the speech data we work with has been transcribed prior to the annotation process, the automatic analysis technique we describe does not use the transcriptions as input. Rather, the speech signal is first processed using basic audio processing techniques in order to extract features from the segments of speech, which are then used for classification using a machine learning model. One might assume that the most straightforward approach would be to use speech recognition technology to transform a speech recording into an automatically obtained transcript and then simply apply a model such as the one developed by Ai and colleagues (2010), which was applied to transcriptions of face to face interactions. However, the state of the art in speech recognition is still too poor to make this a viable option. Although some tutorial dialogue systems such as Scot (Pon-Barry, et al 2006) and ITSPOKE (Forbed-Riley & Litman, 2009) have used speech recognition technology to detect uncertainty in student responses, neither systems required high accuracy of the content in order to make this attribution. For instance, both systems used speech recognition to detect lexical hedges (e.g. I think, I thought, maybe) or pauses to detect uncertainty in student responses. Therefore, despite the great potential value in automatic transactivity or ICC analysis directly from speech, considering the importance of content similarity with prior contributions evidenced in prior work, it remains to be seen what level of accuracy is possible just from the speech signal itself.

The technique we evaluate in this paper is related to prior work on speech processing for other classification tasks. There has been some prior work on automatic assessment of group interactions in the CSCL community focusing on speech as input (DiMicco et al., 2004; Gweon, Kumar, & Rosé, 2009), however that work was more focused on the amount of contribution from each speaker overall rather than anything specific related to the nature of individual contributions. In the language technologies community, some prior work has focused on the nature of conversational contributions, however. For example, Ranganath and colleagues (2009) used acoustic and prosodic features extracted from speech data to predict whether a speaker came across as flirting or not in a speed dating encounter. Similarly, Ang (2002) and Kumar and colleagues

(2006) applied a similar technique to the problem of detecting emotions such as boredom, confusion, or surprise, whereas Liscombe (2005) applied the technique to the problem of detecting student uncertainty. All of this work makes use of signal processing techniques that are able to extract basic acoustic and prosodic features such as variation and average levels of pitch, intensity of speech, amount of silence and duration of speech.

Acoustic and prosodic features are frequently associated with intuitive interpretations that make them an attractive choice to play a role in baseline techniques for these stylistic classification tasks. For example, increased variation in pitch might indicate that the speaker wants to deliver his ideas more clearly. Likewise, volume and duration of speech may signal that the speaker is explaining his ideas in detail, and is presenting his point of view about the subject matter. Such interpretations are grounded in sociolinguistic work related to the way in which speech style specifically (Coupland, 2009; Eckert & Rickford, 2001; Jaffe, 2009) and language style more generally (Fina, Schiffrin, & Bamberg, 2006) reflect both intentional and subconscious aspects of the way in which a speaker positions him or herself within an interaction at multiple levels. These recent accounts build on decades of work beginning with Labov's work on speech characteristics that signal social stratification (Labov, 1966) and Giles' work developing Social Accommodation Theory (Giles, 1984), which describes how speech characteristics shift within an interaction, and how these shifts are interpreted socially. A simplistic interpretation of this work would lead us to believe that hidden within the speech signal are features that enable prediction of social meaning. The Ranganath work (2009) cited above related to detection of flirting supports this view. It is possible to argue that while the essence of transactivity is related to content level distinctions, that it also has a social interpretation, and therefore might be detectable from speech as well. For example, consider that externalizations position students as intellectual leaders within a conversation. However, if true leadership requires that the leader is received as such by the other group members, and transactive contributions indicate that reception, then the occurrence of transactive/ ICC contributions say something about the relationship between speakers. We can then expect that stylistic features that predict positive reception between conversational participants may also predict transactivity/ ICC. The simplest approach to begin such a line of research begins with the types of features used in prior work detecting social aspects of conversations from speech, such as flirting.

Operationalization of the Knowledge Integration Process

When students are working on a given task or a project in a team, they receive a certain amount of information that would help them solve the problem, in the form of a task statement and training materials. In order to solve the given problem, students discuss the materials that are given to them and try to apply them to a potential solution. We are interested in capturing instances when students display reasoning during group discussions that goes beyond what is given and displays some understanding of a causal mechanisms behind the information. Typically some causal mechanism would be referenced in a discussion of how something works or why something is the way it is. In segmenting student talk and identifying which segments display reasoning we are able to quantify amount of reasoning displayed. However, it is important to note that since what we are coding is *attempts* at displayed reasoning, we need to allow for displays of incorrect, incomplete, and incoherent reasoning to count as reasoning. That will necessarily be quite subjective – especially in the case of incoherent explanations. We begin by operationalizing the distinction between non-reasoning statements and reasoning statements, and then we focus on the distinction between reasoning statements that represent new directions within a conversation (i.e., externalizations) from those that build on prior contributions (i.e., ICC).

One important goal in detecting the knowledge integration process is to distinguish instances when students are making their own reasoning explicit from ones that just parrot what they have heard. In our formulation, we consider the task and training materials provided during the experiment to be "given", and we look for contributions where students go beyond that.

Operationalization Step 1: Reasoning Process

Our formulation of what counts as a reasoning display comes from the Weinberger and Fischer's (2006) notion of what counts as an "epistemic unit", where what they look for is a connection between some detail from the given task (which in their case is the object of the case study analyses their students are producing in their studies) with a theoretical concept (which in their work comes from the attribution theory framework, which the students are applying to the case studies). When they have seen enough text that they can see in it mention of a case study detail, a theoretical concept, and a connection between the two, they place a segment boundary. Occasionally, a detail from a case study is described, but not in connection with a theoretical concept. Or, a theoretical concept may be mentioned, but not tied to a case study detail. In these cases, the units of text are considered degenerate, not quite counting as an epistemic unit.

We have adapted the notion of an epistemic unit from Weinberger and Fischer (2006) because the topic of our conversations is very different in nature. The conversations that we analyzed come from a design exercise where 3 participants are asked to design and build an egg holder together. The egg holder will contain an egg, and should protect it from breaking when dropped from a two story high stairwell. As in Weinberger and

Fisher's (2006) notion of "epistemic unit", we look for a connection between two or more concepts. Unlike in Weinberger and Fisher's operationalization of reasoning, where one of the concepts contains at least one detail from the task and the other is a theoretical concept, in our operationalization both concepts can be of either type. We describe our operationalization in detail below. First, examine a segment of a conversation where we have highlighted the instances of displayed reasoning using italics.

s1: *i think we'll need only one rubber band because the rubber band is circular. We can just break it off right*
s3: oh yeah. that's a good idea.
s2: See what are the weights
s1: *it is some significant difference*
s2: *Yeah this is heavier. So this could be on top*
s3: *yeah cause if we did that then that would fall on the bottom, right? It might spin.*

The simple way of thinking about what constitutes a reasoning display is that it has to communicate an expression of some causal mechanism. Often that will come in the form of an explanation, such as X because Y. However, it can be more subtle than that, for example "Increasing the tension makes the spring springier." The basic premise was that a reasoning statement should reflect the process of drawing an inference or conclusion through the use of reason. Note that in the example with the spring, although there is no "because" clause, one could rephrase this in the following way, which does contain a "because" clause: "The spring will be springier because we will increase the tension."

Concepts

The basic building block of a reasoning statement is a concept. We identified 5 types of concepts relevant for our domain, namely theoretical concepts, prior knowledge, physical system properties, emergent system properties, and goals. For each concept, the definition and an example are illustrated in table 1. The examples in the table are from our dataset described in section 3.1, where students are discussing a best approach to build an egg holder. Note that the "system" in this case is the egg holder, plus any materials that are available for use.

Table 1: Definition and examples for the 5 concepts.

Type	Definition	Example
Theoretical concept	principles (i.e. physics principle) and theories	when an object is falling, the force of impact when it hits the ground can be decreased by slowing down the speed.
Prior Knowledge	information based on common sense	Using a small amount of tape would not be enough to hold two bowls together
Physical system properties	elements and characteristics of elements that are available for the system	paper bowl is round, straws are flexible
Emergent system properties	characteristics of elements that appear in a process	stability of an egg holder which emerges as a result of using certain materials
Goal	general believes/ perspectives, anything associated with strong expectations related to points of view	aesthetics of an egg holder, i.e. trying to make the egg holder aesthetically pleasing

Relationship

The presence of multiple concepts in a statement by itself does not determine whether a statement contains reasoning. Rather, the relationship between multiple concepts is the determining factor. For example, a simple list of concepts (e.g., this cup is round, and it is also white) is information sharing, and not reasoning. We identified two types of relationships that signal a reasoning process; 1. Compare & contrast, 2. Cause & effect.

1. Compare and contrast, tradeoff: When the speaker compares two concepts, the speaker is making a judgment, which involves thinking about how two concepts are related to another.
   - The speaker compares two materials ("that" & "rubber band") for his solution: *"I am thinking that might work better than a lot of rubber bands."*

2. Cause and effect: When the speaker uses a cause-and-effect relationship, this process involves establishing the relationship between two concepts through a reasoning process.

The general relation in this category is "doing x helps you achieve y" There are three main types of causal relationship a)cause and effect b)in order to c)analogy. Examples for each of the three types are illustrated below.

- Let's do A because of B: *"Let's use bubble wrap because it cushions the fall"*
- Let's do A in order to achieve B: *"Let's use rubber bands for tying the bag onto the bowl."*
- When a speaker makes an analogy, he is making a link due to the similarity between two concepts. Some of the keywords that signal analogies are "like", "as": *"Oh, you're trying to use the bowl as a parachute."*

Operationalization Step 2: Idea Co-construction (ICC) vs. Externalization

Statements that display reasoning can be either Externalizations, which represent a new direction in the conversation, not building on prior contributions, or ICC contributions, which operate on or build on prior contributions. In our distinction between Externalizations and ICC contributions, we have attempted to take an intuitive approach by determining whether a contribution refers linguistically in some way to a prior statement, such as through the use of a pronoun or deictic expression.

Take the sample conversation we used earlier to illustrate the reasoning contribution. The lines marked with an (E) at the end is a contribution that is categorized as externalization, the ones with a (T) are transactive contributions. The first statement by s1 is an externalization since s1 starts a new topic, thus this contribution is not building on a prior contribution. Subsequent reasoning contributions in this discussion are coded as transactive because they each build on statements that directly precede them.

s1: *i think we'll need only one rubber band because the rubber band is circular. We can just break it off right (E)*
s3: oh yeah. that's a good idea.
s2: See what are the weights
s1: *it is some significant difference (T)*
s2: *Yeah this is heavier. So this could be on top (T)*
s3: *yeah cause if we did that then that would fall on the bottom, right? It might do some spinning. (T)*

Reliability of Annotation

Two coders were initially trained using a manual that describes the above operationalization of reasoning displays and ICC in detail along with an extensive set of examples. After each coding session, the coders discussed disagreements and refined the manual as needed. Most of the disagreements were due to the interpretation of what the students meant rather than the definition of reasoning itself. Therefore, later efforts focused more on defining how much context of a statement could be brought to bear on the interpretation and how. In a final evaluation of reliability for reasoning process, we calculated kappa agreement of 0.67 between two coders over all the data. After calculation of the kappa, disagreements were settled by discussion between the two coders. The coding manual for detecting instances of ICC and externalization is still under development. Our initial round of coding yielded a kappa value of 0.64. The data used in this paper was coded by one coder who has experience with coding for externalization and ICC using a corpus from a different domain.

Automatic Assessment of Reasoning Processes

The purpose of our investigations with speech technology that we report in this paper was to determine the extent to which it is possible to use current machine learning technology paired with simple signal processing preprocessing techniques to distinguish between nonreasoning statements, externalizations, and ICC contributions. We first describe our approach. In the subsequent section, we detail our promising results.

Methods

Our technical approach consists of four main stages: collecting audio data, preparing audio data by transcribing and segmenting the recordings, extracting features from segmented recordings, and displaying predicted scores on a report by applying machine learning. The overall process is detailed in the following subsections.

Collecting the Audio Data

Our corpus was collected in a laboratory setting while students worked face-to-face in groups of three. In this paper, we focus on a subset of the data that has already been collected, transcribed and annotated. The specific task the students are engaged in is to design a contraption to protect an egg when falling the distance of two flights of stairs. This task involves applying a variety of principles of physics. The data we focus on is a 30 minute discussion portion of each 3-student group work session when the participants were designing and

building the egg holder together. In order to collect clean speech with each student on a separate channel, each student wore a directional microphone. Nevertheless, although it is possible to clearly identify the main speaker from an audio file, crosstalk, which is the other participants' voices, could still be heard in the background.

Transcribing and Segmenting the Audio Data

For each audio file, the main thirty-minute discussion sessions were transcribed and manually segmented for further analysis. A total of 8 meetings were collected, transcribed, and segmented according to the following two rules. The resulting data contained a total of 4361 segments.

1. Begin a segment when the main speaker starts talking. If there is silence at the beginning of the file when the main speaker is silent, this means that there will be an "empty" segment in the beginning.
2. A segment should contain the main speaker's continuous speech. If there is an interruption (silence or crosstalk) that lasts for more than 1 second, a new segment should be created. When you create a new segment, there should be two boundaries – one that marks the end of the main speaker's first utterance, and another that marks the start of the next utterance after the pause.

Extract Features from Segmented Recording

After the stage of segmenting the data into units, the next stage involved transforming each segmented unit into a set of feature-value pairs. For the feature set, three types of features were extracted. 1. Acoustic features, 2. Phoneme features, and 3. Auxilary features. All three feature sets reflect "how" the words are spoken rather than the content of the words.

Acoustic features capture certain structural aspects of speech such as amplitude, pitch and energy. More intuitively, these features reflect the intensity and energy level of a given speech segment. For instance, a higher value of amplitude means higher volume of the speaker. If there is variation in the amplitude, this indicates that the speaker's volume varied over time. We collected 4 amplitude features, which are the mean value of amplitude over the whole segment, as well as the mean, median, and variance of the 1 second windows in a given speech segment. Similarly, we extracted 4 pitch and 4 energy features: pitch/ energy of the overall segment, mean, median, and standard deviation of pitch over 1 second windows in a given segment. The pitch features were extracted using the YIN algorithm (De Cheveigné & Kawahara, 2002). In addition to these 12 features, 28 of 40 Mel Frequency Cepstral Coefficients (mfcc) were used in the feature set. The initial 40 mfcc features are the result of applying a set of 40 standard filters, which are available as part of VoiceBox Matlab Toolbox (Voicebox, 2010). The mfccs are standard acoustic features that are commonly used in speech processing. They reflect the distribution of energy level in the given speech. Because using all these 40 features would capture somewhat redundant information, we took the top 28 features using principal component analysis (PCA). The decision to take 28 features was based on a rule of thumb that this number of features is sufficient for a variety of speech classification tasks of a roughly similar nature.

Phoneme related features are based on English phonemes, which are the smallest building block of sound in English that carries linguistic meaning. For instance, the phoneme that distinguishes the words tip and dip are the [t] and the [d] phonemes. Sphinx (CMUSphinx, 2010), a speech recognition system developed at Carnegie Mellon University, identifies 48 phonemes in the English language. Thus, we used the 48 phoneme probabilities as part of our feature set. We believe that using phonemes could capture certain aspects of content that would reflect the coding process used by human annotators or the structure of the language data. For instance, according to our operationalization of a reasoning statement, cause and effect relationships can be used to causally connect two concepts. Certain words, such as "because" or "for" are often used in cause and effect relationships. Therefore, phonemes such as [b] or [f] may occur frequently in statements that contain reasoning contributions. In addition to the phonemes, a phoneme-count feature and phoneme rate were computed. The phoneme-count feature shows the total number of phonemes, which tells us how much the speaker spoke in the given segment. The phoneme rate feature is the number of phonemes divided by length of the segment, and provides us with an estimate of how fast a person spoke.

In addition to the acoustic and phoneme features, three additional features were computed, namely duration of the segment and a speaker feature, and a feature that reflects stylistic language matching. The duration of the segment was the length of the given segment in seconds. The speaker feature was a binary feature, 0 if the speaker of the given segment is same as the speaker of the last segment, 1 otherwise. For the feature that reflects the stylistic language matching, we computed the Kullback-Leibler distance between phoneme probabilities, which is a measure of how different two distributions are from one another.

Once all the features are extracted, we used the Adaptive Boosting machine learning algorithm (Freund & Schapire, 1995) to train a predictive model and then evaluate whether it was possible to automatically assign segments of speech as containing a "non-reasoning/ externalization/ ICC" contribution with high enough accuracy. The Adaptive Boosting algorithm was designed to be resilient to noisy data and outliers because of

the way it trains a model over multiple iterations, and the instances that are misclassified in early iterations receive more attention in the subsequent rounds through a reweighting mechanism.

Displaying Prediction on Report

The ultimate goal of our work is to use the automatic predictions to provide reports to instructors about how collaborative groups are interacting with one another. Thus, once we obtain the frequency of non-reasoning/ externalization/ ICC contributions from a given meeting, we can display the numbers in a graph so that the instructor can get a sense of which groups need support.

Results

Recall that prior to applying machine learning, human annotators manually labeled the data and verified the reliability of the coding process. Next, we used machine learning to produce labels for the data. Table 2 shows how accurate the machine produced labels are compared to the human labels. We achieve an F-score of .56 for distinguishing reasoning from non-reasoning statements. Distinguishing ICC and Externalizations from other statements is lower, at .35 and .32 respectively. Although the recall and precision rates may not seem very high, they are a significant improvement over a baseline (majority class). For all three types of prediction, duration of segment was the top indicator for determining whether a contribution contained reasoning/ ICC or not. In all cases, the length feature was the most important. This result matches the heuristic that if a contribution contains reasoning, it is longer because the speaker needs time to express his thoughts.

When applying machine learning, we took careful steps to avoid the evaluation results being inflated due to overlap in speakers between train and test sets. Namley, we separated the data into two sets, each with a distinct set of students; specifically, a training set for building a model and a test set for testing the accuracy of the model. Given that we had a limited amount of data, we adopted a 10 fold cross validation methodology where we average the performance obtained for each of the ten test sets. For each test set, 1/10 of the data is set apart as test data, and the remaining 9/10 of the data is used to build a model.

Table 2: Machine learning experiment results showing baseline, recall, precision, F-score, and top 3 most predictive features for the prediction of reasoning, ICC, and externalization statements.

Prediction	Baseline F-score	Recall (%)	Precision (%)	F-score (%)	Feature #1	Feature #2	Feature #3
Reasoning vs. other	0.20	0.63	0.51	0.56	Length (27.8%)	Phoneme rate (6.5%)	12th PCA feature (5.2%)
ICC vs. other	0.12	0.72	0.24	0.35	Length (17.8%)	Phoneme 'B' (18.2%)	12th PCA feature (6.7%)
Externalization vs. other	0.08	0.70	0.22	0.32	Length (35.2%)	2nd PCA feature (4.7%)	9th PCA feature (4.3%)

Conclusions and Current Directions

In this paper, we presented our work towards automatic detection of reasoning displays and ICC contributions in speech data. The need for a tool that presents level of ICC contributions has been demonstrated in our previous study where we investigated the needs of instructors who teach project courses (Gweon et al, 2011). The goal of this paper was to develop technology to address such needs. To this end, we have begun with a simple technique, adapted from other stylistic speech classification tasks. Our work shows promise in that 1) humans can distinguish reasoning and non-reasoning statements with acceptable reliability, although reliability on distinguishing externalizations from ICC contributions needs more improvement, and 2) using machine learning, classification of a statement as reasoning/ non-reasoning is feasible, even with limited training data. Results at distinguishing ICC contributions from others are still weak, especially with respect to precision.

In our future work, more sophisticated adaptations of sociolinguistic work might suggest follow-up techniques. Other pieces of work on sociolinguistics of speech style emphasize social interpretations of stylistic shifts within an interaction (Eckert & Rickford, 2001). For example, Social Accommodation Theory (Giles, 1984) emphasizes the important function of stylistic convergences between speakers within an interaction. This work suggests that more complex features computed over patterns of the types of acoustic and prosodic features that we begin with in this paper may be more conducive to high levels of accuracy. In addition to investigating frequency counts and patterns, we also plan to investigate sequencing and timing rather than just quantity as adopted by Kapur and colleagues (2009). In terms of data, we are currently collecting and annotating audio data from additional meetings as well as other contexts to validate our result further as well as testing its generality across a wider variety of student groups. In addition, because the automatic predictions are not perfect, we must also explore how to properly signal instructors about the confidence level of the predictions.

References

Ai, H., Kumar, R., Nguyen, D., Nagasunder, A., & Rosé, C. P. (2010). Exploring the Effectiveness of Social Capabilities and Goal Alignment in Computer Supported Collaborative Learning, *Proc. ITS.* 134-143.

Ang, et. al. (2002). Prosody based automatic detection of annoyance and frustration. In Proc. *International conference spoken language processing,* Denver, Colorado, USA, 2002, pp. 2037– 2039.

Berkowitz, M., & Gibbs, J. (1983). Measuring the developmental features of moral discussion. Merrill-Palmer Quarterly, 29, 399-410.

Chen, M. (2003). Visualizing the pulse of a classroom. In Proc. MM', ACM Press, 555-561.

CMUSphinx Wiki. (2010). http://cmusphinx.sourceforge.net/wiki/

Coupland, N. (2009). Style: Language Variation and Identity: Key Topics in Sociolinguistics, Cambridge University Press.

De Cheveigné, A., & Kawahara, H. (2002). "YIN, a fundamental frequency estimator for speech and music," The Journal of Acoustical Society of America, 111(4), 1917-1930.

De Lisi, R., & Golbeck, S. (1999). Implications of Piagetian Theory for Peer Learning, A. O'Donnell & Alison King (Eds.) *Cognitive Perspectives on Peer Learning*, Lawrence Erlbaum Associates, New Jersey.

DiMicco, J., et. al. . (2004). Influencing group participation with a shared display. *Proc. CSCW*, 614-623.

Eckert, P., & Rickford. J. (2001). Style and Sociolingusitic Variation, Cambridge University Press.

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.

Freund, Y., & Schapire, R. E. (1995). A decision-theoretic generalization of on-line learning and an application to boosting. European Conference on Computational Learning Theory. 23-37.

Fina, A., Schiffrin, D., & Bamberg, M. (2006). *Discourse and Identity*, Cambridge University Press

Forbes-Riley, K., & Litman, D. (2009). Adapting to Student Uncertainty Improves Tutoring Dialogues. *In Proc AIED* 2009: 33-40.

Giles, H. (1984). The Dynamics of Speech Accommodation, Amsterdam: Mouton.

Gweon, G., Kumar, R. Rosé, C. (2009). GRASP: The Group Assessment Platform, *In Proc CSCL*.

Gweon, G., et. al. (2011). A Framework for Assessment of Student Project Groups Online and Offline. Puntambekar & Hmelo-Silver (Eds.) Analyzing Interactions in CSCL, Springer. 293-317.

Jaffe, A. (2009). Stance: Sociolinguistic Perspectives, Oxford University Press.

Joshi, M., & Rosé, C. P. (2007). Using Transactivity in Conversation Summarization in Educational Dialog. Proceedings of the SLaTE Workshop on Speech and Language Technology in Education

Kapur, M., & Kinzer, C. K., (2009). Productive failure in CSCL groups. *In Proc, CSCL 2009*. 4: 21-46.

Kay, J.,et al. (2006). *Wattle Tree: What'll It Tell Us?*, University of Sydney Technical Report 582, January 06'.

Kim, T. Chang, et. al. (2008). Meeting Mediator: Enhancing Group Collaboration with Sociometric Feedback, *In Proc. of CSCW,* San Diego, CA, 457-466.

Kumar, R., Rosé, C. P., & Litman, D. (2006). Identification of Confusion and Surprise in Spoken Dialogusing Prosodic Features , *Proceedings of Interspeech 2006*.

Labov, W. (1966). The social stratification of English in New York City, Washington DC: Center for Applied Linguistics.

Liscombe, J. Venditti, J, & Hirschberg, J. (2005). Detecting certainness in spoken tutorial dialogues. *In Proc. Interspeech 2005*. 1837-1840.

Pianesi, F., et al. (2008). Multimodal support to group dynamics. *Personal and Ubiquitous Computing.* 12(3), 181-195.

Pon-Barry., et al. (2006). Responding to Student Uncertainty in Spoken Tutorial Dialogue Systems. IJAIED 16(2)

Ranganath, R., Jurafsky, D., & McFarland, D. (2009). It's Not You, it's Me: Detecting Flirting and its Misperception in Speed-Dates. *Proceedings of EMNLP 2009.* 334-342.

Rosé, C. P., et al. (2008). Analyzing Collaborative Learning Processes Automatically: Exploiting the Advances of Computational Linguistics in Computer-Supported Collaborative Learning , IJCSCL. 237-271

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

Teasley, S. D. (1997). Talking about reasoning: How important is the peer in peer collaboration? In L. B. Resnick (Eds.), Discourse, tools and reasoning: Essays on situated cognition, 361-384.

Voicebox (2010). http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html

Weinberger A., & Fischer F. (2006). A framework to analyze argumentative knowledge construction in computer supported collaborative learning. Computers & Education; 46, 71 – 95.

Acknowledgments

This research was supported in part by NSF EEC grant number 064848.