Automatic classification of activities in classroom discourse

Zuowei Wang*, Xingyu Pan, Kevin F. Miller, Kai S. Cortina

Combined Program in Education and Psychology, University of Michigan, 610E University Ave, Rm1406, Ann Arbor, MI 48109, USA

ARTICLE INFO

Article history:
Received 6 March 2014
Received in revised form 27 May 2014
Accepted 31 May 2014
Available online 9 June 2014

Keywords:
Improving classroom teaching
Pedagogical issues
Teaching/learning strategies

ABSTRACT

Classroom discourse is the primary medium through which teaching and learning occur. Managed skillfully, it can provide an opportunity for students to develop their understanding and to profit from the ideas of their peers and the teacher. Yet it is difficult for teachers to be mindful of the nature and distribution of classroom discourse at the same time as they juggle other instructional concerns. It is possible to record, transcribe, and analyze classroom discourse, but it is not possible to do this quickly enough to give a teacher timely feedback. We report on the development and validation of an automated system for recording and analyzing aspects of classroom discourse that can result in timely feedback. Based on the LENA system, it aims to identify three common discourse activities: teacher lecturing, whole class discussion and student group work. The system consists of a speech processing module (diarisation performed by the LENA system) and an activity detection module that detects the discourse activities by using classification analysis. Results showed that our automatic detection of discourse activities converged well with those of human coders. The system enables timely and relatively inexpensive generation of a classroom discourse profile, which helps teachers to visualize and potentially improve their classroom discourse management skills.

© 2014 Elsevier Ltd. All rights reserved.

1. Introduction

Student discussion can provide learners the opportunity to work through their understanding and learn from ideas of others. Successful classroom discussions require skillful management, and there is evidence that this is a challenge for many teachers. We will report on a project aimed at developing and validating an automated system that addresses one of these challenges – the lack of awareness teachers have about the distribution of talk during lessons they teach. The goal of this project is to determine whether we can

(a) discriminate between different classroom discourse structures
(b) using automated models of speech data
(c) rapidly and with accuracy levels comparable to that of human coders.

Before describing our model and its validation, we will first discuss, (1) why classroom discourse is important to learning, (2) research using trained human coders to identify classroom discourse structures and the relation of discourse to learning, and (3) what kinds of data we might expect an automated system to be able to identify from classroom discourse.

1.1. Why classroom discourse is important, and why it's elusive

The centrality of discourse to teaching and learning has long been recognized. In their Professional Standards for Teaching Mathematics, the National Council of Teachers of Mathematics (NCTM) states that “the teacher of mathematics should promote classroom discourse in...”

* This study was supported by a grant from the National Center for Educational Research (NCER) at the Institute of Educational Science (R305A100178, PI: Kevin Miller; Making Room for Student Thinking: Using Automated Feedback, Video-Based Professional Development, and Evidence-Based Practice Recommendations to Improve Mathematical Discussion).

* Corresponding author. 610 E. University Ave, Rm1400J, Ann Arbor, MI, 48109, USA. Tel.: +1 734 615 8887.

E-mail addresses: zwwang@umich.edu (Z. Wang), xypan@umich.edu (X. Pan), kevinmil@umich.edu (K.F. Miller), schnabel@umich.edu (K.S. Cortina).

http://dx.doi.org/10.1016/j.compedu.2014.05.010
0360-1315/© 2014 Elsevier Ltd. All rights reserved.
which students listen to, respond to and question the teacher and one another” (NCTM, 1991). Their revised standards (NCTM, 2000) also stress the importance of communication in mathematics classes: instructional programs should enable children to “organize and consolidate their mathematical thinking through communication”, to “communicate their mathematical thinking coherently and clearly to peers, teachers and others” (NCTM, 2000).

These recommendations are consistent with research on the relation between classroom discourse and student learning. For example, Smith (1977) has discovered that classroom discourse that allows more student involvement leads to more critical thinking and better learning outcome. Tobin (1984, 1986) reported that better questioning practice of the teacher that encouraged student involvement promoted learning. To facilitate teachers’ questioning, Hiebert and Wearne (1993) designed lessons that emphasized the relationship between different mathematical concepts rather than just practicing computation procedures. They found teachers teaching these lessons asked students more questions and these students performed better than their peers in exams. On the other hand, bad management of classroom discourse could be harmful to student learning. Gürler (2013) examined the classroom discourse and students’ understanding of the concepts in a calculus class. He found that students’ confusions were closely related with the instructor’s failure to attend to concept shifts in the discourse.

Benefits of classroom discussion have also been identified in reading classes. In a comprehensive review that covered efforts aiming to promote classroom discourse in reading classes since the 1960s, Murphy, Wilkinson, Soter, Hennessey, and Alexander (2009) found that most of these projects successfully increased student talk and decreased teacher talk, and they improved students literal and inferential comprehension.

Although the importance of classroom discourse is widely acknowledged, observations indicate that most teachers are not proficient in promoting student discussion. Pianta, Belsky, Houts and Morrison (2007) reported their analysis of the observation of over 10,000 classrooms across 10 sites in the U.S. They found that most teachers only provided their students with feedback on the correctness of their answer, rather than asking them to elaborate on their reasoning. This followed the simple teacher initiation–student response–teacher evaluation IRE pattern (Mehan, 1979, p. 80), reflecting a transmissionist view of learning that has been criticized by many contemporary researchers (e.g. Inagaki, Hatano, & Morita, 1998; Nassaji & Wells, 2000; Waring, 2009). For example, Inagaki et al. (1998) argued that the IRE instruction left little room for “negotiation”; in contrast, if a teacher allowed other students to elaborate or criticize their original responses, they would have the opportunity to construct their mathematical thinking by assimilating similar ideas from their peers or revising their current conceptual model to accommodate conflicting ideas from others, both of which have been confirmed by their study.

This pattern of limited student opportunity to explain mathematics is not universal. Sims (2008) compared the amount of mathematical statements made by students in American and Chinese classrooms. She found that in the US students on average made only 21% of mathematical statements and questions, whereas in China the number was 69%. The fact that US students were not given enough opportunities to express their mathematical thinking in class may indicate a lack of classroom discourse facilitation from the teachers.

One obstacle that prevents teachers from gaining expertise in discourse management is the implicit nature of this skill and the ephemeral nature of classroom discourse. It is difficult to pay attention to something like the distribution of talk when one is teaching, at the same time that teachers need to worry about managing the learning of their students and the accuracy of their own explanations. Traditional teacher training does not provide many opportunities for teachers in training to develop skill at managing student discussions. Eilam and Poyas (2006) described this process as consisting of three elements: learning theoretical knowledge, observing experienced teachers and teaching practice. None of these are directly aimed at improving skill in managing classroom discourse. Rather, it has been assumed that novice teachers can gain this skill automatically through experience.

1.2. Research efforts to decipher classroom discourse

A number of efforts have been made to increase teachers’ awareness in their classroom discourse. For example, Carlisle, Kelcey, Berebitsky, and Phelps (2011) observed comprehension lessons for a year and they focused on three dimensions of reading instruction: pedagogical structure, teacher-directed instruction, and support for student learning. They found teachers’ propensity in these activities predicted students’ reading outcome. Teacher training programs have been carefully designed to help teachers better organize classroom talk in mathematics classes (e.g. Chapin, O’Connor, & Anderson, 2009) as well as reading classes (Murphy et al., 2009, for a comprehensive review). Researchers have also been investigating factors that affect the quality of classroom discourse. Using transcripts of lessons, Bellack and his colleagues discovered some universal features of talk moves in different classrooms, based on which they categorized teaching into four categories: structuring, soliciting, responding and reacting (Bellack, Kliebard, Hyman, & Smith, 1966). The dynamic change of teaching activities among the four categories has been further used to define the “teaching cycle”, which reflects the instruction features in a class. For example, the average length of teaching cycles could represent the pace of instruction; by analyzing the initiator of each teaching cycle (be it teacher or students), one could also obtain the “relative proportion of teacher and pupil discourse” (Kliebard, 1966).

In an effort to help teachers visualize their classroom instruction, Walsh (2006) identified four classroom “modes”, including Managerial, Materials, Skills and systems and Classroom context. These categories were then introduced to teachers to help them perform “self-evaluation of teacher talk” (SETT), in which teachers watched their own classroom recordings and identified their teaching activities based on the given categories. Final interviews with these teachers indicated that their awareness of discourse management was improved after the practice.

Cazden and Beck (2003) summarized five discourse features that can be consciously controlled by a teacher: speaking rights and listening responsibilities, teacher questions, teacher feedback, pace and sequence, and classroom routines. Variations on these features result in different types of classroom discourse, which further influences students’ learning. For example, a teacher can encourage student involvement by giving them more “speaking rights” and making sure other students take their “listening responsibilities”; she can also slow down the “pace” by providing more wait time before calling a student to answer a question so that other students may have longer time to think about it.

In addition to these studies of classroom discourse in general, there are several discourse coding systems focused specifically on mathematical discourse. For example, the Mathematical Quality of Instruction (MQI) provides specific standards on a variety of perspectives...
of instruction such as teachers' interactions with students and students' participation (Hill et al., 2008). By comparing a recorded class to the MQI standards, researchers will be able to identify teachers' strength and weakness and help them make progress accordingly.

The video project (Stigler, Gonzales, Kawanaka, Knoll, & Serrano, 1999) conducted as part of the original TIMSS (Trends in International Mathematics and Science Study) project (Peak, 1996) is perhaps the most ambitious effort to date to code classroom discourse in classrooms internationally. In the initial round of research, the team recorded middle school mathematics lessons in the U.S., Japan, and Germany, transcribing all classroom discourse. In later work, the method has been extended to additional countries and expanded to include science as well as mathematics. The code book used by this project distinguished between whether or not interactions were public (i.e., involving the whole class as a unit) or individual/small-group work, and whether the teacher or the student was providing the bulk of the information (Stigler, Gallimore, & Hiebert, 2000). Based on the coding system, eighth-grade mathematics lessons from the U.S., Germany and Japan have been compared and significant cross-country differences have been identified. For example, in the classes of Germany and Japan much more topics were “developed” instead of simply “stated” by the teacher, whereas in the U.S. the pattern was flipped (Stigler et al., 2000).

These coding systems have yielded interesting and informative findings in educational research. However, they have rarely been used to provide feedback to teachers as a result of two practical limitations. Coders need to receive many hours of training and certification before they can begin to use the instruments, and coding is itself a time-consuming activity. This makes it very difficult for teachers to get timely feedback on their lessons, which may lessen the impact of the delayed feedback and teachers' ability to use it to modify their practice. In addition, most of these training projects took months of work from teachers (Murphy et al., 2009); this also greatly limits broader implementation.

1.3. Automated modeling of recorded speech

If we could automatically distinguish between whether or not the teacher or student is speaking, and record both the length of individual utterances and the total amount of talk during a lesson, it might be possible to automatically identify meaningful activity patterns during lessons based on the nature of discourse involved. Such a system would not be able to evaluate the quality of what was said, but information about the quantity and distribution of classroom talk might nonetheless provide a useful index to some important aspects of a lesson.

One successful effort to obtain timely automated feedback related to children's language is the Language Environment Analysis system (LENA; Ford, Baer, Xu, Yapanel, & Gray, 2008). The LENA system was designed to record the language environment of very young children and to produce measures of the quality of language produced by and received by the wearer. This measurement has been applied in a number of studies that investigated how children's language environment affected their cognitive development in different contexts. For example, Weisleder and Fernald (2013) recorded all—day interactions of children and their caregiver with LENA and found that infants who experienced more child-directed speech showed faster language development.

In this project, we adapted the LENA to a different purpose, testing whether we could use it in classrooms, worn by the teacher, to identify meaningful aspects of classroom discourse. Our usage of LENA was different from its original design in several ways: firstly, it was designed for preschool children whereas we used it in elementary school; secondly, instead of asking a child to wear the recorder, we asked the teacher to wear it; thirdly, the classroom environment may be noisier and certainly has more speakers than does the home environment. Despite these differences, we found that the LENA can work in the classroom environment (Wang, Miller, & Cortina, 2013). In that study, we found that LENA can reliably recognize the identity of the speaker in a recorded class. For example, it can distinguish whether the speaker is the teacher or a student (although it cannot differentiate different students); it can also detect whether multiple speakers are speaking at the same time. When no one is speaking, LENA identifies that as silence. We were also able to modify the LENA software to improve its ability to differentiate student from teacher talk.

The goal of this project is simply to determine whether we can use the data produced by the LENA system to produce and validate a model that can replicate human coders' identification of classroom activity structures. Specifically, we explore the question whether we can develop a classification model that can distinguish three classroom activities (lecturing, whole class discussion and student group work) in mathematics lessons based on LENA data.

2. Material and method

2.1. Participants

Thirteen elementary school mathematics lessons from 12 teachers were recorded. Students in these lessons varied from 1st to 4th grade. All the classrooms were recruited in southeast Michigan.

2.2. Speech recognition system

We collected speech data by asking teachers to wear the Language Environment Analysis (LENA) system, developed by the LENA Foundation (Ford et al., 2008) while they taught their regular mathematics lessons. The LENA system identified segments of speech into the following categories as coming from: 1) adults (male or female) classified as near to or far from the device, 2) children also classified as near or far, 3) overlapping speech from multiple speakers, 4) noise, or 5) broadcast speech as from a TV or radio. It also reported the average volume of each utterance.

1 Based on unsystematic piloting, we believe that the current LENA system will work with students up to about age 10. For older students, some student voices are identified as adult speech.

2 Because it assumes a child wears the device, the software distinguishes between a ‘key’ child and other children; this distinction was not useful for us.
2.3. Procedure

Teachers were asked to turn on the LENA recorder, put it into a pouch worn around their neck, and teach their mathematics lesson as they normally would. After the lesson, they were asked to turn off the device and connect it to a laptop computer, which extracted the recording data and sent it to our remote server. On our server, their recording files were first analyzed by the LENA program. In some cases, teachers neglected to turn off the LENA at the end of the lesson. These instances could be identified based on length of lesson, and we determined the actual end of the lesson from the recording.

Based on our previous study (Wang et al., 2013) we adjusted the LENA results to produce the categories we were interested and to better fit the classroom environment. After these procedures, for each recorded class we had a time stamped file that contained information about speaker identification and signal intensity (volume): we identified each utterance as 1) teacher talk, 2) student talk, 3) overlapping speech, and 4) non-speech (silence or noise).

2.4. Human coding

The recorded speech was first divided into 30-s segments. Then two coders independently coded each segment based on which of the following activity types dominated: 1) teacher lecturing, 2) whole class discussion and 3) student group work. In teacher lecturing, the teacher introduces new learning content to students, explains tasks or giving extended feedback to students without student response—the teacher dominates the classroom discourse. In whole class discussion, the teacher and students have conversations about the learning content and this conversation happens on the whole class level. The discussion should be accessible to the majority of students in class. In group work, students work in groups and talk with each other thus making the classroom noisy, and can only hear their partners and the conversations are different among the groups. These three categories are the outcomes of the classification (three potential classification results). Due to the complex speech environment of the classroom, not every segment could be categorized with certainty, so along with the classroom activity we also assigned a confidence rating ranging from 1 to 3 to each categorized fragment. A confidence level of 3 indicates very confident, 2 means considerably confident and 1 shows a lack of confidence.

2.5. Classification model

Using the LENA data, we first calculated for each 30 s segment the following parameters from the time stamped speaker identification file: the percentage of teacher talk time, the percentage of student talk time, the percentage of overlapping talking time, the percentage of silence time, and the average volume. These parameters were the data provided to the classification algorithm as the independent variables.

We used the data from one of our coders (henceforth “T” for Trainer) to develop our classification model, which was then tested on the data from the other (henceforth “V” for Validator).

The classification analysis used the Random Forest (RF) algorithm (Breiman, 2001). RF consists of three steps: first, a number of sample training datasets (N) are generated from the original data by bootstrapping; second, for each sample dataset, a classification tree is formed producing a total of N trees which form a “forest”; third, the classification results from all trees are aggregated by majority voting, providing the final classification results.

Our training set consisted of all selected segments coded by coder T with a confidence rating of 3 (high confidence). We then used the “randomForest” package from R (Liaw & Wiener, 2002) to fit the parameters obtained from our automated system to the human judgments. After training, the model was applied to the classification of all the segments (no matter the level of confidence), and the resulting classifications were first compared to coder T and then to coder V’s independent coding.

It should be noted that to test the robustness of the classification model we did not have any assumption about how different characteristics about specific classrooms or schools could affect the classification. In other words, no hierarchical linear modeling (HLM) was used in the current study. Using an HLM method may lead to better classification results but it will limit the finding to specific classrooms or schools, which does not fit the purpose of the current study.

3. Results

3.1. Descriptive data of the manual coding

The average duration of all the recordings was 44 min, ranging from 31 to 62 min. In total, we had 608 min of class recordings, thus 1216 30 s episodes to categorize. We had two coders identified as T(trainer) and V(validator). We asked coders to assign a confidence level for each segment. For coder T, this broke down as follows: 62% confidence level 3 - high confidence; 30% confidence level 2; 8% confidence level 1. For coder V the corresponding percentages were 66%, 22% and 12%, respectively. Table 1 provides a breakdown of the coding of segments for the training sample (the high-confidence segments from Coder T) as well as the two full samples from both coders.

Inter-rater reliability is shown using a confusion matrix in Table 2. Entries on the diagonal represent percentage of agreement for both coders across the three classroom teacher categories. For Lecturing, 42.2% of segments were put in this category by both coders; Coder T

<table>
<thead>
<tr>
<th></th>
<th>Lecture</th>
<th>Discussion</th>
<th>Group work</th>
</tr>
</thead>
<tbody>
<tr>
<td>Training sample</td>
<td>42.98%</td>
<td>32.74%</td>
<td>24.25%</td>
</tr>
<tr>
<td>Full sample (T)</td>
<td>47.04%</td>
<td>34.49%</td>
<td>18.45%</td>
</tr>
<tr>
<td>Full sample (V)</td>
<td>50.62%</td>
<td>32.75%</td>
<td>16.62%</td>
</tr>
</tbody>
</table>

Table 1: Distribution of categories in training and testing samples.
identified an additional 4.6% as lecturing, for an overall agreement of 94%. For the Discussion category, agreement was at 75%, and 78% for Group Work. Adding all the three percentages on the diagonal together, inter-rater reliability is 83%. Cohen's Kappa = 0.72, \( p < .001 \), with 95% confidence interval of (0.68, 0.76), indicating substantial agreement (Landis & Koch, 1977).

3.2. Differentiating classroom activities using current parameters

Table 3 shows the “center” values (means) of the five prediction parameters (independent variables) for the three classroom activities. It can be seen that the Lecture activity is characterized by a high proportion of teacher talk; in comparison, Discussion contains even distribution of teacher talk, student talk and silence. In Group Work, there is a lot of overlap. MANOVA analysis revealed that the values of the five parameters differ significantly across the three classroom activities, \( F (10, 2394) = 181.71, \ p < .0001 \), suggesting classification is promising.

3.3. Reliability of the automatic classification algorithm

The reliability of the classification algorithm is evaluated first on the training dataset then on two cross-validation datasets. Firstly, the comparison between the model prediction and T’s manual coding illustrates the training performance, that is, how well the algorithm fits the training dataset. Table 4 illustrates the training performance of the classification. The classification accuracy for Lecture, Discussion, and Group Work is 0.881, 0.797, and 0.833, respectively, which result in an overall classification accuracy of 84.37%. The Kappa between the model prediction and the training data is 0.76 with a 95% CI of (0.72, 0.80).

The first cross-validation is performed by comparing the model prediction with coder T's complete coding data. Since the model training was only based on a subset of T's codes, this comparison illustrates how well the model performs on unlearned data. Comparing the algorithm prediction with T's manual coding, the classification accuracy for lecture, discussion, and group work is 89.3%, 80.6%, and 84.7% respectively, with an overall accuracy of 85.2% and a Cohen's Kappa = 0.77, with a 95% CI of (0.73, 0.80) (Table 5).

The second cross-validation is performed by comparing the model prediction to coder V's manual coding.3 Because the model was built upon coder T's coding, who coded independently from coder V, the model prediction is independent with coder V's manual coding. Comparing the algorithm and V's coding, the classification accuracy for Lecture, Discussion, and Group Work is 93.1%, 74.5%, and 76.3%, respectively, with an overall accuracy of 83.0% and a Cohen's Kappa = 0.73, with a 95% CI of (0.69, 0.76) (Table 6). It can be noted that both the overall accuracy and the Kappa value in this independent cross-validation are comparable to those between the two human coders. Thus the classification model is indistinguishable from a real human coder.

Finally, to evaluate the model in a potential application, we compared the total amount of lecturing, whole class discussion and group work calculated from the coding of the two coders and that from the model. Figs. 1–3 showed the scatter plots of these comparisons. Table 7 summarized the correlation coefficients.

4. Discussion

In this study we explored the possibility of using speech processing technology to detect three activities of classroom discourse: teacher lecturing, whole class discussion and student group work. We first parsed the lesson into 30 s bins, and asked a human coder to identify each segment as lecture, whole class discussion, or group work. We then developed a model for those segments the human coder classified with high confidence. The detection was achieved by a classification model that used the randomForest algorithm to model the human coding.

To determine the validity of the classification model, we compared the model–coder consistency with the coder–coder consistency. Cohen's Kappa between our two coders is 0.72, with a 95% confidence interval of (0.68, 0.76). According to Landis and Koch (1977), this value indicates substantial agreement. In comparison, Cohen's Kappa between the model classification and the validation coder is 0.73, with a 95% confidence interval of (0.69, 0.76). The model learned from the first coder (Trainer) and then its results are compared to the other coder (Validator), whose coding was not used in the model development. It should be noted that no statistically significant difference is found between the consistency of the model and the validation coder and the inter-coder consistency, since the 95% confidence intervals of the Kappa's overlap. This indicates high validity of the currently classification model.

The selection of the three activities to categorize classroom discourse is based on previous analysis of classroom activities (Stigler et al., 2000). In their coding system, they classified classroom activity based on two dimensions: whether the interaction is public or private, or whether the teacher or the student is providing the bulk of information. Based on the two dimensions, we derived three common classroom activities: teacher lecturing (where the teacher is providing the bulk of information and the interaction is public), whole class discussion

---

3 As a test, we swapped the “Trainer” and “Validator” and retrained and retested the model, and the results were similar.
The validity of this categorization approach is supported by the agreement of coding: the amount of each category is similar between the two coders, most of the episodes are classified with considerable confidence by the two coders (88% and 83% with a confidence level of 2 or 3 from the two coders respectively), and the confusion matrix (Table 2) shows high convergence. We retained this simple categorization approach hoping that this study can serve as a test bed for the broader application of speech recognition technology to the classroom.

The classification model provides the basis for creating meaningful feedback for teachers’ classroom discourse management. For example, the classification model allows the calculation of the amount of discussion in a recorded class. As indicated by Fig. 2 and Table 7, the total amount of whole class discussion obtained from the model and that from manual coding showed high convergence; thus our model can automatically provide teachers with reliable feedback on this measure.

Table 3
Mean and SD (in parentheses) of parameters for each classroom activities.

<table>
<thead>
<tr>
<th></th>
<th>Student%</th>
<th>Teacher%</th>
<th>Overlap%</th>
<th>Silence%</th>
<th>Signal-level</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lecture</td>
<td>6.0 (7.3)</td>
<td>60.2 (16.8)</td>
<td>5.8 (10.0)</td>
<td>18.8 (14.9)</td>
<td>–560.66 (151.79)</td>
</tr>
<tr>
<td>Discussion</td>
<td>17.6 (15.3)</td>
<td>26.2 (20.2)</td>
<td>6.8 (10.6)</td>
<td>24.2 (20.5)</td>
<td>–678.08 (170.79)</td>
</tr>
<tr>
<td>Group work</td>
<td>8.6 (12.8)</td>
<td>32.7 (16.5)</td>
<td>45.7 (24.1)</td>
<td>4.3 (10.1)</td>
<td>–413.08 (213.58)</td>
</tr>
</tbody>
</table>

Table 4
Confusion matrix for model training (number of 30 s segments).

<table>
<thead>
<tr>
<th></th>
<th>Model</th>
<th>Lecture</th>
<th>Discussion</th>
<th>Group work</th>
<th>Classification error</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lecture</td>
<td>282</td>
<td>29</td>
<td>8</td>
<td>11.59%</td>
<td></td>
</tr>
<tr>
<td>Discussion</td>
<td>31</td>
<td>194</td>
<td>18</td>
<td>20.16%</td>
<td></td>
</tr>
<tr>
<td>Group work</td>
<td>2</td>
<td>28</td>
<td>150</td>
<td>16.67%</td>
<td></td>
</tr>
</tbody>
</table>

Table 5
Confusion matrix, model prediction vs. T’s manual coding (number of 30 s segments).

<table>
<thead>
<tr>
<th></th>
<th>Lecture</th>
<th>Discussion</th>
<th>Group work</th>
<th>Classification error</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lecture</td>
<td>467</td>
<td>53</td>
<td>3</td>
<td>10.71%</td>
</tr>
<tr>
<td>Discussion</td>
<td>75</td>
<td>348</td>
<td>9</td>
<td>19.45%</td>
</tr>
<tr>
<td>Group work</td>
<td>24</td>
<td>14</td>
<td>210</td>
<td>15.33%</td>
</tr>
</tbody>
</table>

Table 6
Confusion table, model prediction vs. V’s manual coding (number of 30 s segments).

<table>
<thead>
<tr>
<th></th>
<th>Lecture</th>
<th>Discussion</th>
<th>Group work</th>
<th>Classification error</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lecture</td>
<td>485</td>
<td>35</td>
<td>1</td>
<td>6.91%</td>
</tr>
<tr>
<td>Discussion</td>
<td>99</td>
<td>321</td>
<td>11</td>
<td>25.53%</td>
</tr>
<tr>
<td>Group work</td>
<td>22</td>
<td>36</td>
<td>187</td>
<td>23.68%</td>
</tr>
</tbody>
</table>

(where students are providing the bulk of information and the interaction is public) and group work (where the interaction is private, and students are providing the information).

The validity of this categorization approach is supported by the agreement of coding: the amount of each category is similar between the two coders, most of the episodes are classified with considerable confidence by the two coders (88% and 83% with a confidence level of 2 or 3 from the two coders respectively), and the confusion matrix (Table 2) shows high convergence. We retained this simple categorization approach hoping that this study can serve as a test bed for the broader application of speech recognition technology to the classroom.

The classification model provides the basis for creating meaningful feedback for teachers’ classroom discourse management. For example, the classification model allows the calculation of the amount of discussion in a recorded class. As indicated by Fig. 2 and Table 7, the total amount of whole class discussion obtained from the model and that from manual coding showed high convergence; thus our model can automatically provide teachers with reliable feedback on this measure.

Fig. 1. Total amount of lecturing (in percentage of the whole lesson) calculated from human coding and the classification model.
The automated coding is substantially cheaper than human coding, but its real advantage is the timeliness of coding made possible with the automated system. It is reasonable to assume that feedback that is given before the next time you teach is more useful than feedback given later, when memories of the class in question have been overwritten with later experiences. It is not feasible to give such timely feedback with human coding on any reasonable scale, but it is possible with the LENA-based system. In ongoing work, we are providing teachers with nightly email messages giving them feedback on that day's lessons, which is available to them before they teach again.

Given the importance of student discussion (NCTM, 2000, 1991; Smith, 1977; Tobin, 1984, 1986) and the lack of student involvement in class (e.g., Pianta et al., 2007), providing teachers with feedback on how much their students were involved in discussion could potentially help the teachers improve the facilitation of discussion.

In addition to identifying overall levels of different kinds of discourse activity, the automated system can also be useful in identifying segments of a lesson for further analysis. Although it can identify sections of lesson where whole class discussion occurs, the automated system is not able to look at the content or the quality of discussion. But it can identify sections of a lesson for further analysis, where one could look at the quality of discussion or the clarity of teacher explanations during lecturing. This combination of automated and human coding may yield an efficient chimera that could provide a quick evaluation of the quality of a lesson.

The analysis reported here does not look at the flow of activity within a lesson, although work by Givvin and her colleagues (Givvin, Hiebert, Jacobs, Hollingsworth, & Gallimore, 2005) shows that there are distinctive patterns of lesson activity flow that are characteristic of different countries. The relative ease and inexpensiveness of automated coding will facilitate the identification of such patterns in the flow of lessons, as well as identifying distinctive approaches to teaching whose effectiveness can be explored.

The automatic classification of classroom activities is limited to mathematics lessons. As pointed out by Carlisle et al. (2011), instruction is domain specific and is limited by social and organizational classroom factors. Thus the classification model here may not work in other subjects such as reading. The system is also limited to elementary school classes due to the LENA speech processing engine—as mentioned earlier in this paper, children who are older than fourth grade are often recognized as the teacher, which leads to inaccurate classification results.

Fig. 2. Total amount of discussion (in percentage of the whole lesson) calculated from human coding and the classification model.

Fig. 3. Total amount of group work (in percentage of the whole lesson) calculated from human coding and the classification model.
The clearest limitation of the method employed here is its inability to engage with the content of speech. Due to this limitation, the method used here should not be used in high-stake evaluations of teachers' instruction. In a recent experimental study, Strong, Gargani and Hacifazlioglu (2011) have found that even human judges who are quite consistent with their evaluation of teachers often fail to identify instructional practices that are related to better student achievement. Similarly, the current system, though being very consistent, may miss important instructional practices that are directly related to student achievement. Luckily, it is not an inherent limitation of automated systems for analyzing classroom discourse. In the same way that methods such as latent semantic analysis have proven to be effective at identifying the quality of written text (Landauer, Foltz, & Laham, 1998), techniques such as word-spotting (Barnwal, Sahni, Singh, & Raj, 2012) that look for instances of particularly important keywords hold promise for adding information about the quality of discourse to the methods used here.

5. Conclusion

In conclusion, in this study we successfully build an automatic classification model to detect three kinds of classroom activities. The model converges well with human coders and the reliability is comparable to human coders, as reflected by Cohen's Kappa. The automatic detection of different classroom activities sheds light on a new approach of classroom discourse research. Moreover, the model can be used to create quick and meaningful feedback for teachers to improve their skills in classroom discourse management.

References


