Toward an Automatic Speech Classifier for the Teacher

Bahar Shahrokhian Ghahfarokhi(✉), Avinash Sivaraman, and Kurt VanLehn

Arizona State University, Tempe, AZ, USA
{bshahrok, asivara6, kvanlehn}@asu.edu

Abstract. Our system classifies audio from microphones worn by the teacher in order to determine (1) whether the teacher is addressing the whole class or talking to individuals or groups of students. In the latter case, it determines (2) whether the teacher is giving formative feedback, giving corrective feedback, chatting socially, or addressing administrative or workflow concerns. This paper reports the initial accuracy of this system against human coding of middle school math classroom behavior. We also compared audio collected through professional hardware versus more accessible alternatives.

Keywords: Intelligent tutoring system · Educational data mining · Multimodal learning analytics

1 Introduction

"Classroom orchestration" refers to the teacher's management of the activities, students and information in classes that integrate small group work, individual work, and whole-class work [1]. Orchestration systems are intended to help teachers and students achieve productive interactions and learning in such classrooms. A key problem addressed by orchestration systems is increasing the teacher's awareness of the state of all the students and their interactions [2]. Awareness is particularly difficult to maintain when students are working in small groups, because much of their interaction is spoken and inaccessible to the teacher. Thus, we are focusing on analysing classroom speech when the main class activity is working in small groups.

In order to focus the research while maintaining its generality, we are exploring just two categorization schemes. Each has been investigated in prior research and seem generally useful for helping teachers maintain awareness.

The first classification scheme divides the teachers' activity into (1) addressing the whole class, (2) talking with students, (3) talking with experimenters and (4) not talking. This classification is useful for several purposes. First, an orchestration system should supress all alerts when the teacher talking to the whole class because teachers would not be able to attend to them. It might send only high priority alerts when the teacher is talking to students, since those conversations are punctuated by times when the teacher is waiting for the students to answer and thus has a limited capacity to attend to alerts. Prior work [3, 4] has focused on similar categories, but included a

I. I. Bittencourt et al. (Eds.): AIED 2020, LNAI 12164, pp. 279–284, 2020.
https://doi.org/10.1007/978-3-030-52240-7_51

category for lecturing and omitted the category for talking to experimenters. Lecturing did not occur in our math classrooms.

The second classification scheme divides teacher's conversation with students into formative interaction, corrective interaction, and several non-instruction categories. The distinction between formative and corrective instruction is traditional and goes by many names. If the teacher points to incorrect work, gives strong hints about correctness or explains how to do correct work, then their instruction is called corrective, didactic or teaching by telling. If teachers elicit explanations from students, encourage them to think more or pose challenging questions without answering them, then their instruction is called formative, teaching by eliciting or formative assessment [5–10]. In the classes we observed, the teachers were all attempting to given formative instruction exclusively. However, we often observed that their conversations with students were corrective. An orchestration system could collect such episodes and present them to the teacher as part of a post-class debriefing sessions. The system might even give teachers feedback on their conversations in the middle of class whenever teacher had some spare time and wanted to get such feedback.

Like many of our predecessors [3, 11, 12, 15], we analysed the speech with acoustic features only. We did not attempt to convert the speech to text and use lexical features, semantic features, or other natural language processing. On the one hand, this probably reduces the accuracy of the classification, particularly for the distinction between formative vs. corrective teacher feedback. On the other hand, this study take place in a live classroom while preserving the privacy of students. It also increases the chance that system is domain general. That is, after it has been trained on classes engaged in one set of lessons and tasks, it can be used without change on a wider set of lessons and tasks. However, testing the generality of our system remains as future work. Here we report its accuracy against human coders.

2 Data Collection and Analysis

2.1 Raw Data Collection

The raw data for this paper was gathered during a class trial of FACT in spring of 2019. This trial consisted of 6, 50-min periods of 8th grade students working on specific sets of mathematics lessons, called Classroom Challenges [10] using our FACT web-based platform [13, 14]. Each period consisted of 9 to 16 groups of mostly two students, but we only had permission to collect multimodal data from 31 groups, out of which we annotated 20 groups (64% of the data), due to reasons like low audio quality or class being too short. 40.4% of all annotated students are male.

During this study, the teachers wore a lavalier microphone and carried a tablet to access FACT and manage classroom lessons and students. While working in a group, students each used their own tablet to access a shared group workspace, typically with their heads down over the tablets while they talked with each other. Most students wore an Audio-Technica PRO 8HEx headset microphone connected to a Tascam DR-40 digital audio interface. Students in 2 groups per classroom wore throat microphones.

In each session 4 groups' interactions were captured using a video camera with its own shotgun mic and a second channel for a boundary mic laid on the group's table. A shoulder-mounted video camera followed the teacher and recorded the lavalier mic's output. Most tablets' screens were also captured by screen recorders. All these recordings were used by the human coders but not by the system. Thus, the human coders had much more information than the system, as befits a gold standard.

2.2 Human Segmentation and Coding

We used ELAN [16] to synchronize all the media files and code them. Although several Elan tiers were used, only two are relevant here.

The Teacher View tier was for classifying the teacher's behaviour into 'Whole Class', 'Admin' and 'Group Interaction'. The 'Whole class' label indicated that the teacher is giving a whole class announcement. This was usually done to explain the activity or the user interface to the students before starting the activity. The Admin category indicated that the teacher was talking to the one the FACT system admins. The Group Interaction label indicated that the teacher was talking to a group or student. Most of the class activity was group work, but this label was used even when students were working individually. In this experiment, teachers never lectured and were rarely silent. This coding created segments of arbitrary length.

We then divided up the 'Group Interaction' segments, which tended to be many minutes long. Using a Teacher Group Interaction tier, we created 30 s segments, because prior work showed that varying the length of segments did not greatly impact accuracy [3, 16]. Human annotators labelled the segments as 'Formative', 'Corrective', 'Overhead', 'Workflow' or 'Chat' coding. A segment was labelled 'Formative' or teaching by eliciting, if for example the teachers instead of giving explanations and feedback, keep students engaged in solving problems, which requires the teachers to analyze the students' work, detect the line of reasoning being followed and then ask questions that push the students further along that line [15]. A segment was labelled 'Corrective' or teaching by telling, if the teacher's instruction is more direct. The 'Overhead' label indicates that the teacher is talking to the individual group about any issues with FACT. The 'Workflow' label indicates that the teacher was discussing the class or tasks (e.g., moving from one task to another) but not giving instruction.

Two human coders labeled 15% of the all the 30-s segments. Inter- rater agreement was considered acceptable with Cohen's kappa K = 0.75. For the segments that two coders disagreed with each other, that segment was discussed until both coders agreed on one label, if the agreement was not achieved that segment was not considered for the next step.

2.3 Audio Processing

To process the audio, we initially removed noise using audacity [17]. We extracted acoustic features using openSmile [18]. We created time series features using the tfresh [19] Python package. Since the number of extracted features was quite large, we also used different feature selection algorithms like PCA and Pairwise correlations to reduce the redundancy of the feature set.

3 Result

This section reports the accuracy of our system against the human annotations. Classifiers were machine-learned using random forest, deep learning (forward only, not RNN), KNN, decision tree, additive logistic regression and SVM. Accuracy was measured with 10-fold classification. Random forest yielded the best result for all analyses.

- *Teacher's View* tier: Categorization into 'Whole Class', 'Group Interaction' and 'Admin'.
  - Teacher's audio captured via tablet: Accuracy 0.88, Kappa 0.84
  - Teacher's audio captured via lavalier microphone: Accuracy 0.91, Kappa 0.88.
- *Teacher Group Interaction* tier: Categorization into 'Formative', 'Corrective', 'Overhead', 'Workflow' or 'Chat'
  - Teacher's audio captured via tablet: Accuracy 0.77, Kappa 0.65
  - Teacher's audio captured via lavalier microphone: Accuracy 0.74, Kappa 0.61.

For recording devices, it appears that the tablet's audio was nearly as good as wearing a lavalier mic.

4 Discussion and Comparison to Prior Work

To the best of our knowledge this is one the first studies working on automatic speech classifier for math teachers in middle school classrooms. To evaluate the absolution (as opposed to relative) accuracy, it helps to compare our results to prior work, here we focused on studies analyzing teachers' speech in live classrooms. We must mention that, due to difference in experiment settings, features type, approach, purpose and data annotation, direct comparison of the result is not possible.

Wang et al. [4] trained a classifier to automatically label 30 s segments of teacher's speech, recorded via LENA system, into 'teacher lecturing', 'whole-class' discussion and 'student group work'. Their overall classification accuracy of 84.37% is comparable with our overall classification accuracy of 83.34%.

D'Mello et al. [3, 20–23] also found comparable accuracies to ours. They used linguistic as well as acoustic features of the teacher's speech, and they studied language arts classrooms instead of math.

A challenge for this and similar projects is protecting the privacy of students' speech. Although we only extracted non-lexical features from audio, the database has the full speech audio. Perhaps we can extract the features and discard the audio recording before the end of class to better preserve the students' privacy.

To increase the performance of our classes, other than adding lexical features, we are also planning to extract more features from other multi modal inputs such as our ITS logs, teacher's position, etc.

Acknowledgements. Supported by NSF FW-HTF 1840051.

References

1. Dillenbourg, P., Nussbaum, M., Dimitriadis, Y., Roschelle, J.: Design for classroom orchestration. Comput. Educ. **69**, 485–492 (2013)
2. Prieto, L.P., Holenko Dlab, M., Gutiérrez, I., Abdulwahed, M., Balid, W.: Orchestrating technology enhanced learning: a literature review and a conceptual framework. Int. J. Technol. Enhanc. Learn. **3**(6), 583 (2011)
3. Blanchard, N., D'Mello, S., Olney, A.M., Nystrand, M.: Automatic classification of question & answer discourse segments from teacher's speech in classrooms. International Educational Data Mining Society (2015)
4. Wang, Z., Pan, X., Miller, K.F., Cortina, K.S.: Automatic classification of activities in classroom discourse. Comput. Educ. **78**, 115–123 (2014)
5. Wiliam, D., Lee, C., Harrison, C., Black, P.: Teachers developing assessment for learning: impact on student achievement. Assess. Educ.: Princ. Policy Pract. **11**(1), 49–65 (2004)
6. Cauley, K.M., McMillan, J.H.: Formative assessment techniques to support student motivation and achievement. Clear. House: J. Educ. Strateg. Issues Ideas **83**(1), 1–6 (2010)
7. McMillan, J.H.: Formative Classroom Assessment: Theory into Practice. Teachers College Press, New York (2007)
8. McMillan, J.H.: Classroom assessment. principles and practices for effective instruction. Allyn & Bacon, A Viacom Company, 160 Gould St. (1997). www.abacon.com. Needham Heights, MA 02194
9. Black, P., Wiliam, D.: Assessment and classroom learning. Assess. Educ.: Princ. Policy Pract. **5**(1), 7–74 (1998)
10. Burkhardt, H., Schoenfeld, A.: Assessment in the service of learning: challenges and opportunities or Plus ça Change, Plus c'est la même Chose. ZDM **50**(4), 571–585 (2018)
11. Martinez-Maldonado, R., Dimitriadis, Y., Martinez-Monés, A., Kay, J., Yacef, K.: Capturing and analyzing verbal and physical collaborative learning interactions at an enriched interactive tabletop. Int. J. Comput.-Support. Collab. Learn. **8**(4), 455–485 (2013)
12. Gweon, G., Jain, M., McDonough, J., Raj, B., Rosé, C.P.: Measuring prevalence of other-oriented transactive contributions using an automated measure of speech style accommodation. Int. J. Comput.-Support. Collab. Learn. **8**(2), 245–265 (2013)
13. FACT. http://fact.engineering.asu.edu/. Accessed 27 Feb 2020
14. http://fact.asu.edu/. Accessed Feb 2020
15. Herman, J.L., et al.: The implementation and effects of the mathematics design collaborative (MDC): early findings from Kentucky ninth-grade algebra 1 courses. CRESST Report 845. National Center for Research on Evaluation, Standards, and Student Testing (CRESST) (2015)
16. Martinez, R., Wallace, J.R., Kay, J., Yacef, K.: Modelling and identifying collaborative situations in a collocated multi-display groupware setting. In: Biswas, G., Bull, S., Kay, J., Mitrovic, A. (eds.) AIED 2011. LNCS (LNAI), vol. 6738, pp. 196–204. Springer, Heidelberg (2011). https://doi.org/10.1007/978-3-642-21869-9_27
17. Audacity Team: Audacity (R): free audio editor and recorder [computer program]. Version 2.1.0 (2014). https://sourceforge.net/projects/audacity/. Accessed 27 Feb 2020
18. Eyben, F., Wöllmer, M., Schuller, B.: OpenSmile: the Munich versatile and fast open-source audio feature extractor. In: Proceedings of the 18th ACM International Conference on Multimedia, pp. 1459–1462 (2010)
19. Christ, M., Braun, N., Neuffer, J., Kempa-Liehr, A.W.: Time series feature extraction on basis of scalable hypothesis tests (tsfresh–a python package). Neurocomputing **307**, 72–77 (2018)

20. Donnelly, P.J., et al.: Automatic teacher modeling from live classroom audio. In: Proceedings of the 2016 Conference on User Modeling Adaptation and Personalization, pp. 45–53 (2016)
21. D'Mello, S.K., et al.: Multimodal capture of teacher-student interactions for automated dialogic analysis in live classrooms. In: Proceedings of the 2015 ACM on International Conference on Multimodal Interaction, pp. 557–566 (2015)
22. Blanchard, N., et al.: Semi-automatic detection of teacher questions from human-transcripts of audio in live classrooms. International Educational Data Mining Society (2016)
23. Jensen, E., et al.: Toward automated feedback on teacher discourse to enhance teacher learning. In: Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems, pp. 1–13 (2020)