Multi-Sensor Modeling of Teacher Instructional Segments in Live Classrooms

Patrick J. Donnelly[1], Nathaniel Blanchard[1], Borhan Samei[2], Andrew M. Olney[2], Xiaoyi Sun[3], Brooke Ward[3], Sean Kelly[4], Martin Nystrand[3], and Sidney K. D'Mello[1]
[1]University of Notre Dame, USA; [2]University of Memphis, USA;
[3]University of Wisconsin-Madison, USA; [4]University of Pittsburgh, USA
118 Haggar Hall, Notre Dame, IN, 46556, USA
pdonnel4@nd.edu

ABSTRACT
We investigate multi-sensor modeling of teachers' instructional segments (e.g., lecture, group work) from audio recordings collected in 56 classes from eight teachers across five middle schools. Our approach fuses two sensors: a unidirectional microphone for teacher audio and a pressure zone microphone for general classroom audio. We segment and analyze the audio streams with respect to discourse timing, linguistic, and paralinguistic features. We train supervised classifiers to identify the five instructional segments that collectively comprised a majority of the data, achieving teacher-independent F_1 scores ranging from 0.49 to 0.60. With respect to individual segments, the individual sensor models and the fused model were on par for Question & Answer and Procedures & Directions segments. For Supervised Seatwork, Small Group Work, and Lecture segments, the classroom model outperformed both the teacher and fusion models. Across all segments, a multi-sensor approach led to an average 8% improvement over the state of the art approach that only analyzed teacher audio. We discuss implications of our findings for the emerging field of multimodal learning analytics.

Categories and Subject Descriptors
Computing methodologies~Discourse, dialogue and pragmatics • Computing methodologies~Supervised learning by classification • Social and professional topics~K-12 education

General Terms
Experimentation, Human Factors

Keywords
classroom discourse; dialogic instruction; speech recognition; automatic feedback; educational data mining

1. INTRODUCTION
There are many complexities to our educational systems, but few would dispute that the teacher plays a central role in improving student outcomes. In turn, improving educational outcomes requires, in part, improving teacher instruction. This raises the following central question: What are teachers doing and how can we make what they do better?

Research has shown that classroom instruction continues to be dominated by traditional instructional techniques, such as lecture, recitation, and seatwork [13], despite the fact that there are more appealing alternatives [13, 23]. For example, dialogic instruction is a form of classroom discourse that is characterized by thought-provoking discussions between teachers and students with the goal of facilitating a meaningful exchange of ideas intended to elicit deeper thought and analysis. Several large scale studies has shown that the dialogic approach to classroom instruction positively correlates with student engagement [14] and achievement [2, 19], yet its implementation in classrooms is scarce [13].

Research has also demonstrated that the quality of classroom instruction can be enhanced with teacher training programs [4]. For example, research has demonstrated that discussing data-driven analysis of classroom practices with teachers correlates with student achievement [15]. The ability to provide teachers with qualitative and formative feedback on their instruction is paramount to improving and refining their teaching strategies over time. This typically requires that teacher instruction techniques are assessed by classroom observations so that appropriate feedback can be provided [13]. Regrettably, current efforts to assess teacher instruction rely on manual coding by trained observers, a labor and cost intensive endeavor that cannot be deployed practically, broadly, nor uniformly.

In an attempt to address this critical bottleneck, our study is part of a large multi-disciplinary project in the emerging field of multimodal learning analytics [6]. The idea is to automatically analyze classroom instructional practices towards the goal of providing quantitative and actionable information to researchers, teachers, teacher educators, and professional development personnel. As a step in this direction, we present an approach to automatically identify key instructional segments (Question & Answer, Procedures & Directions, Supervised Seatwork, Small Group Work, and Lecture) in live classrooms. We focus on the identification of instructional segments because they represent an intermediate and quantifiable level of analysis of a teacher's use of class time. The use of these segments can ostensibly be compared between teachers regardless of personal teaching styles, subject matter, academic curriculum, or other factors.

Our approach involves multi-sensor recordings of teacher and classroom audio that are subsequently analyzed at the linguistic, paralinguistic, and discourse levels. We focus on audio recordings because they incur far fewer privacy concerns compared to video recordings and their utility has previously been demonstrated for this task [3, 22, 28]. Furthermore, audio is a key component

Conference'10, Month 1–2, 2010, City, State, Country.

modality for multimodal interactions, so there is an opportunity to make a more basic contribution to multimodal learning analytics.

2. RELATED WORK

There is a long research history on the use of audio (and video) to study instructional practices and student behaviors in live classrooms [1, 10] - most notably see the Measures of Teacher Effectiveness (MET) project [9]. However, the recordings are typically processed by humans; automatic analyses of classroom video and audio are rare. Thus, while there is an active field of automatic student or learner modeling [26], the complementary field of teacher modeling is just beginning to emerge.

In the initial attempt at automatic identification of components of instructional segments from audio recordings, Wang et al. [27, 28] recorded 608 hours of classroom audio from 12 teachers in 1st to 4th grade mathematics classes. They divided the audio into 30 second segments. Two trained coders annotated each segment with respect to the dominant classroom activity: teacher lecture, class discussion, or group work and provided a level of confidence for their annotations. Working independently, the coders achieved an agreement of 83%. From the audio recordings, the authors extracted features describing timing patterns of teacher speech, student speech, or silence. They trained a random forest classifier to identify the dominant class activity of each 30 second segment, reporting an overall accuracy of 84% when compared to the human annotations.

Although this result is an important first step in automated teacher activity analysis, some methodological concerns are warranted. In particular, the same audio segments appeared in both the training and testing sets, albeit using annotations from two different observers (inter-rater reliability of 83%). Second, the authors did not validate their model independent of the teacher, permitting instances from each teacher to appear in both the training and test sets. Third, all coding was completed offline solely based on the audio recording, thereby losing important visual contextual cues that would be available during a live-coding session. Finally, the authors considered only three types of classroom activity, seemingly forcing each 30 second segment into one of these broad categories and perhaps overlooking more subtle differences (e.g., individual work vs. group work).

A second recent study attempted to automatically differentiate between teacher-centric (teacher directions) and student-centric activities (group work) [22]. The authors simulated four different class sessions in their lab with students from a nearby primary school. Monitoring the teacher-facilitator, the authors extracted 144 features from five modalities: eye-tracking, EEG, accelerometer, video, and audio. Following the class sessions, a researcher segmented the session into 10 second windows and annotated the type of activity performed in each window. Using a random forest classifier, the authors achieved 63% accuracy in differentiating between the teacher explaining tasks to the whole class vs. monitoring students working in groups. This study was the first to consider eye-tracking, EEG, and accelerometer features for this task, although the authors found these modalities to be less successful than audio or video. In fact, they found that audio features alone were the most useful as they achieved 56% accuracy using audio alone. Unfortunately, the potential generalizability of these findings are limited by the small dataset of four simulated class sessions in a lab, covering only a single subject in which all sessions followed a common lesson plan.

We previously explored automatic detection of Question & Answer segments on a dataset of 21 class sessions obtained from three teachers [3]. Using only recordings of teacher audio, we extracted 11 features pertaining to the timing of speech and rest patterns and achieved an overall accuracy of 67% (AU-ROC of 0.78) based on teacher-independent validation.

In a subsequent study, we recorded teacher audio from 76 classroom recordings collected from 11 teachers [7]. We extracted timing patterns of speech and rest, features derived from automatic transcriptions of the teacher's speech, and low-level acoustic features. We trained supervised machine learning models to identify occurrences of five key instructional segments, validated independently of teacher. We were able to detect the instructional segments above chance levels, with target segment F_1 scores ranging from 0.45 to 0.55. Additionally, we compared different temporal segmentations of the recordings (window sizes ranging from 30 seconds to five minutes) and found no single window size ideal across the different segments. Lastly, we compared the three feature modalities and found different feature types useful for different segments, a finding we explore in the present work.

3. CONTRIBUTION AND CHALLENGES

We describe a low-cost, non-invasive, multi-sensor approach to automatically analyze teacher instructional activities in live class sessions. The present study is an extension of our previous work [7] that exclusively focused on a single sensor (a teacher mic). Here, we augment our approach with a second sensor that records general classroom activity, including audio from both the teacher and students. We consider seven different types of features, including those modeling the timing of teacher and student speech, automatic transcriptions of the teacher's speech, and acoustic features derived from both the teacher and classroom recordings. We then train supervised classification models to identify five different key instructional segments, validated independent of the teacher in order to generalize to new teachers. Lastly, we combine the output of individual feature set classifiers sets into a multi-sensor late fusion model to learn the optimal weights of each feature type. An overview of our approach is shown in Figure 1.

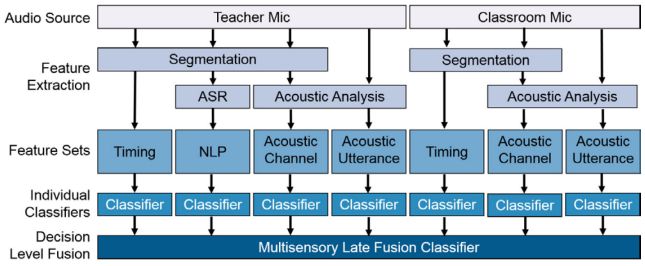

Figure 1. Overview of the system

The research is novel in the following ways. First, unlike previous work that focused on ecological data sets, but with a single sensor, [7, 28] or utilizes multiple sensors, but on a small non-ecological data set [22], we focus on multi-sensor modeling with a large ecologically-valid data set. Second, our sensor combination is unique in that it fuses a high-fidelity teacher mic with a low-fidelity classroom mic, thereby posing interesting challenges on how to jointly analyze audio from the two. Third, comparisons among the two sensors allows us to ascertain if there are any added advantages to incorporating a classroom mic (present

study) to previous state of the art results involving the teacher mic only [7]. This has important practical implications due to the comparative ease of proceeding with a single sensor solution.

A key challenge of our work is that we are attempting to solve a difficult inference problem - identifying teacher instructional strategies at a fine-grained level - entirely from audio. A further complication is that the ground truth is derived from live observations of classes where the human coders used visual information, to contextualize the coding. For example, the coder may observe that students are working on a task in pairs rather than individually, an assessment that may be difficult to determine from the audio recording alone, especially if there is not much talking.

Our emphasis on audio is motivated by privacy concerns with recording video in real-world classrooms and the high-cost of physiological sensing (as discussed in [6]). Despite using audio alone, speech is considered to be a key component modality for multimodal interactions [21] and we analyze the audio stream(s) at multiple levels (linguistic, paralinguistic, discourse via timing patterns), each arguably representing a different mode of communication. If successful, then future work can consider incorporating additional sensors and modalities, such as skeletal tracking with Kinect's or using web-cams after de-identifying individual students.

4. DATA COLLECTION

Data was collected from U.S. middle school literature, language arts, and civics classes. Over the course of two semesters, we collected data from 56 class sessions, covering eight different teachers (one male, seven female) across five schools. The teachers were not coached in any way and were asked to carry out their normal lesson plan, allowing the capture of an unbiased real-world sample of teachers' instructional practices.

Each class session lasted between 30 minutes to 90 minutes, depending on the school, with an average class length of 66:41 (mm:ss). The recordings, totaling over 62 hours, capture the gamut of events typical in a classroom, from focused instruction to distracting interruptions.

We recorded two channels of audio (see Figure 2): one of the teacher and the second capturing classroom speech encompassing both the teacher and the students. Each teacher wore a wireless microphone to capture their speech. Based on previous work [6], a Samson 77 Airline wireless microphone was chosen for its portability, noise-canceling abilities, and low-cost. The teacher's speech was captured and saved as a 16 kHz, 16-bit single channel audio file.

Although it may be useful to record individual students, it is impractical because of concerns for privacy and equipment cost. Therefore, we placed a second microphone in the classroom (usually on the blackboard) in order to capture general classroom speech, including student audio. Based on experiments in classroom environments [6], we selected the Crown PZM-30D pressure-zone microphone (PZM). PZMs are omnidirectional boundary microphones that are placed on large surfaces (e.g., on the teacher's desk) and help to minimize phase interference between direct and reflected sound. This classroom microphone captures both teacher and student speech, however, unlike the teacher recording, the fidelity of the classroom channel is not sufficient to automatically transcribe student speech.

Figure 2. Samson 77 Airline Microphone (left) and Crown PZM-30D (right)

4.1 Coding Instructional Segments

The Nystrand and Gamoran classroom coding scheme [18] considers a hierarchy of classroom events, ranging from general to more specific. In this work, we focus on their coding scheme for 17 possible teacher instructional activities (e.g., Lecture, Discussion). An observer trained in the coding scheme was present during each recorded class session. The observer used software developed for live coding of classroom discourse to annotate instructional segments as they occurred [18].

There were two trained observers in this study. Each class session was "live" coded by one observer, whose coding was subsequently verified by a second observer offline. Disagreements were discussed and the coding refined until both observers reached complete agreement. The annotated instructional form the "ground truth" used in our classification models.

4.2 Analysis of Instructional Segments

We focus on detecting the five most frequently occurring segments in the data: Question & Answer (19%), Procedures & Directions (23%), Supervised Seatwork (14%), Small Group Work (10%), and Lecture (9%). Ironically, Discussion, an instructional segment important to student achievement [14], represented only 1% of the dataset. Since Discussion is related to Question & Answer (both feature whole-class, interactive discourse), we combined the two segments in this study, leading to a Question & Answer occurrence of 20%.

There were 11 additional types of instructional segments that occurred less frequently, such as an occasional distraction, the discipline of a student, a test or quiz, or students reading silently. Individually, these segments were rare, but together they comprised 25% of the data. Although we did not build models for these segments, we retain them as a Miscellaneous category.

We refer to [18] for a full description of each of the five key segments. Briefly, in a *Question & Answer* segment, the teacher asks a question, one or more students may respond, and the teacher evaluates the response. In *Procedures and Directions*, the teacher mainly communicates instructions, often as a transition to another instructional segment. *Small Group Work* divides the class into groups of two or more students to collaborate on a task. During *Supervised Seatwork* segments, students work independently on tasks while the teacher walks around and answers individual questions that arise. *Lectures* involve the delivery of pre-scripted material and may include films and other media.

The individual teachers divide time differently, a reflection of their unique styles. There was also considerable variation over the different class sessions, even within each teacher, which reflects differences in daily lesson plans. In general, an individual teacher or class session may not contain all five instructional segments.

5. MODEL BUILDING

5.1 Partitioning Audio into Windows

A system that is to be deployed in classrooms will not have the benefit of human coders present and will be unable to determine the boundaries of instructional segments. Although we could potentially build detectors to automatically infer segment boundaries, this itself is a separate research problem not investigated here. Instead, we divided each class recording into consecutive non-overlapping temporal windows for classification. We examined non-overlapping windows so as to not bias our results through classification of particularly easy or difficult segments multiple times in the dataset.

Each window was assigned a label using the segment annotations provided by the classroom coders. This label corresponds to the ground-truth for training and validation of the models. For the cases in which a particular window spanned more than one annotation, the dominant classroom activity (in terms of time) was chosen as the segment annotation. An example of this process for a 60-second window is illustrated in Figure 3.

The average segment was 2.9 minutes long, which we used to inform our selection of window sizes. Specifically, we explored windows sizes ranging from 30 to 180 seconds, which allowed us to explore the tradeoff between large and small windows - large windows afforded more data for analysis but yielded fewer instances and increased the likelihood of multiple segment annotations per window; vice versa for shorter windows.

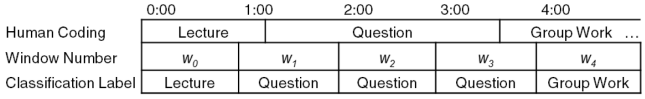

Figure 3. Example of the windowing scheme for a sample of five minutes of class time considering a 60 second window

5.2 Utterance Segmentation

Teacher Channel: Each recording represents an uninterrupted channel of teacher audio lasting the duration of the class session. We first divided the audio signal into smaller audio chunks, each of which ideally represents an utterance spoken by the teacher. Because patterns of speech and rest differ between teachers, as do unintentional noises such as breathing or coughing, we employed a general method to segment utterances [6] that avoided overfitting to specific teachers, thereby increasing generalizability to new teachers.

For this task, we used our approach described and validated in [6], and briefly reviewed here. First, we analyzed the amplitude envelope of the audio to identify moments of silence in which the amplitude of the signal dropped below a predefined noise threshold [3]. We used this silence as a breakpoint from which to partition the recording into *potential teacher utterances*. Next, we processed this set of potential utterances with the Microsoft Bing automatic speech recognition (ASR) system. Potential utterances discarded by Bing were considered to be false alarms. We retained those that contained speech and discarded the others (e.g., background noise, heavy breathing).

Applying this process on our dataset of 56 classroom recordings yielded 32,134 potential utterances, 18,662 (58%) of which were retained as containing speech. The average length of these speech utterances was 5.51 seconds (SD = 8.45), however 2% of the utterances lasted over thirty seconds in length; for example, when the teacher makes a long statement without pausing.

Classroom Microphone: Although ASR of student's speech would be ideal, the single classroom microphone did not have sufficient fidelity for this purpose and it was not possible to mic individual students. Therefore we aimed to segment student utterances from the classroom audio stream in order to ascertain high-level discourse markers, such as the occurrence and duration of student speech.

Because the classroom microphone was placed near the front of the classroom, it captured both teacher and student speech unlike the teacher's headset microphone that captured only the teacher's speech. For this reason, the aforementioned approach used to segment the teacher's audio is insufficient to identify student utterances. We addressed this issue with an alternative approach discussed and validated in [6] and briefly described below.

First, we generated *potential student utterances* using the Azure[1] speech recognition API applied to the classroom channel. The Azure ASR system created a set of transcribed words, each time-stamped individually. From this set, we reconstructed utterances by combining consecutive words unbroken by moments of silence lasting more than one second. The classroom channel was sufficiently noisy so we only used the transcriptions to discriminate between speech and non-speech.

The resulting set of potential student utterances contains both teacher and student utterances because the classroom microphone recorded the entire classroom. Next, we used the segmentation of the high-fidelity audio from the teacher's microphone to filter out the teacher's speech from the set of potential student utterances. The process assumed that occurrences when the teacher and students speaking simultaneously was rare and when it did occur, we assumed the teacher's speech takes priority. In total, we identified 18,123 utterances containing student speech.

5.3 Feature Extraction

We extracted features from each of the windows (see Section 5.1 above) to create the instances used to train and test our classification models. We explored seven different groupings of features, which we refer to as features sets. Four of these sets were computed from the teacher's microphone, adapted from [7], and the other three sets are derived from the classroom microphone.

5.3.1 Teacher Channel

Utterance Timing Features: For each partitioned window of time, we identified any utterances present within the window. If any utterance straddled the boundary of the partitioned window, only the portion of the utterance contained within the window was considered. Using the timing of the utterances and considering gaps between adjacent utterances as rest, we constructed sequences of speech and rest. We then extracted six features from the speech component of the speech-rest sequences: the number of occurrences, the total length of all utterances, the mean and standard deviation of utterance duration, and the durations of the longest and shortest utterance. We extracted the same six features from the timing of the rest patterns. We added one more feature representing the normalized temporal position of the window

[1] The Azure speech recognition API is available online: https://azure.microsoft.com/en-us/marketplace/partners/speechapis/

proportionate to the total length of the classroom recording, resulting in a total of 13 features.

Natural Language Features: We selected the Bing ASR for use in this study based on prior experiments [7] and its ability to freely transcribe large volumes of audio, an important consideration for broader deployment. We employed a natural language processing (NLP) tagger [20] that was specifically designed to classify questions and has been used in studies of classroom discourse [24, 25]. We considered a set of high-level NLP features because the topics covered vary between class sessions and a bag-of-words analysis may not generalize across teachers and class sessions. We extracted 37 natural language features, including the presence/absence of parts of speech (e.g., adjectives, nouns) and particular terms (e.g., *what, how, why*) per utterance. Because the ASR transcriptions are time-stamped at the utterance level rather than the word level, we analyzed the entire utterance even if it overlapped with the time window. We calculated the sum and mean of the 37 natural language features in each window, yielding a total of 74 values.

Acoustic Features by Channel: We extracted a set of acoustic features using the Music Information Retrieval toolbox for Matlab [16]. These features were not extracted from segmented teacher utterances but directly from the entire window of teacher audio. They include descriptors that characterize volume, spectra, and the frequency curve of the signal. We used the following features: six statistical moments describing the spectral distribution (*centroid, flatness, spread, skew, kurtosis*, and *entropy*); *brightness*, a measure of high energy (above 1500 Hz); *zero crossing*, a measure of noisiness counting the times the signal changes sign; two measures of *roll-off*, the frequency cutoff such that 85% and 95% of the total energy was below the cutoff; *root-mean-square energy*, a global measure of the energy of the signal; *low energy*, the proportion of 50 millisecond frames with below average energy; and 13 *Mel-frequency cepstral coefficients,* a cepstral representation of the power spectrum. We supplemented these 25 features with additional measures of voiced frequencies [8], including the *global mean* frequency and *standard deviation* of all voiced frequencies, the *number of blocks* of voiced syllables, and the *average* and *standard deviation* of these blocks. In all, we extracted 30 acoustic features from each window.

Acoustic Features by Utterance: We also extracted acoustic features based on our segmentation of teacher utterances, which contain speech only. We calculated the acoustic features for each utterance contained within the window and averaged these values across all utterances. These features differed from our aforementioned acoustic features because they consider only the parts of the audio stream we previously determined to contain teacher speech (see Section 5.2), excluding moments of silence. Therefore, we excluded the five features describing the voiced frequencies (described above) because the signal was limited to individual spoken utterances rather than the whole channel. We included the 25 descriptors of volume and spectra of the signal described in the previous section.

5.3.2 Classroom Channel

Utterance Timing Features: Considering both the teacher and student utterances derived from the segmentation of the classroom mic (see Section 5.2), this feature set examined teacher:student speech patterns. For the teacher, we calculated the number of utterances, the sum, the mean, standard deviation, minimum, and maximum of the duration of the utterances in the window. We augmented these with the same six features derived from the student utterances. Lastly, we examined the number of alternating turns between teacher and students, resulting in 13 features.

Acoustic Features by Channel*:* We also calculated the same set of 30 features as in Section 5.3.1 based on the classroom microphone using the entire audio stream in each window.

Acoustic Features from Utterances*:* Finally, we calculated 25 acoustic features from the classroom microphone based on our segmentation described in Section 5.2. These features were identical to those calculated for the teacher's utterances (see Section 5.3.1) but were derived from the student utterances.

5.4 Supervised Classification

We generated a total of 210 features for each window. These features were used to train supervised classification models to identify instructional segments. As noted in Section 4.2, there was considerable data imbalance due to an infrequent occurrence of some segments and the high variance in use between different teachers and across class sessions. Therefore, we prioritized the five most common segments, and trained an individual binary classifier to differentiate each segment from all others. For example, the Lecture classifier determines if each instance in the dataset is an example of a Lecture segment vs. one of the other four target segments or one of the 11 infrequently occurring Miscellaneous segments. We considered binary models for each instructional segment type rather than a multi-class model in order to facilitate comparisons of the different features sets for each segment and to ascertain optimal window sizes per segment type.

We considered the Naïve Bayes classifier using the WEKA machine learning toolbox [12]. Naïve Bayes was chosen based on preliminary experiments with several other standard classifiers (e.g., logistic regression, support vector machine, *k*-nearest neighbor, decision tree, random forest), for comparison with our previous work [7], and because of its popular and successful use as a classifier in many domains [17].

Because the five instructional segments of interest individually represent only 9-23% of our dataset, we supplemented our *training* data with additional synthetic instances generated by the Synthetic Minority Over-sampling (SMOTE) algorithm [5]. SMOTE increases the number of instances of the underrepresented minority class label (target label) to eliminate skew in the training set. Oversampling was applied to the training set only; the class distributions in the testing set were not altered.

6. EXPERIMENTS AND RESULTS

All experiments were conducted with leave-one-teacher-out cross validation. For each of the eight teachers, all instances stemming from that teacher's class sessions were added to the test set and the training set was formed from instances of the other seven teachers. This process was repeated for each teacher, and the results were calculated from a confusion matrix aggregated across teachers. This approach allows better generalization to new teachers by preventing the models from overfitting based on characteristics of individual teachers.

In terms of metrics, accuracy, or recognition rate, is not an ideal measure when there is class imbalance as they are in our data. Therefore, we evaluated the efficacy of our binary (target segment vs. all others) models by examining the F_1 score, a balance of precision and recall, for the target segment (e.g., Question & Answer). This ensured that we focus on the model's ability to detect the segments of interest, which was always the minority,

rather than prioritizing the dominant class label (i.e. the other category).

In previous work [7], we compared window sizes and observed that no window size was ideal for all instructional segments. Therefore, we report the best window size for each segment.

6.1 Comparison of Channels

We first compared each of the seven features sets to ascertain if one set or sensor could suffice or if a multi-sensor approach was warranted. These results are displayed in Table 1 for the teacher mic (rows 1-4) and classroom mic (rows 6-8). We first note that no individual feature set consistently outperformed the others, suggesting the importance of a multi-level analysis of the data. With respect to comparing the best feature set(s) across sensors, the best teacher and best classroom feature sets were only marginally different for Question & Answers (0.61 vs. 0.60), Procedures & Directions (0.49 vs. 0.48), and Lecture (0.49 vs. 0.50). However, we observe improvements for Supervised Seatwork (0.46 to 0.55) and Small Group Work (0.48 to 0.55) when considering the best classroom feature set compared to the best teacher feature set.

We also created fusion models for each instructional segment: a teacher model and a classroom model, shown in Table 1 as rows 5 and 9, respectively. The teacher model contained four features, the individual outputs for the four feature set classifiers pertaining to the teacher mic. The classroom model contained three features, the outputs of the three feature set classifiers pertaining to the classroom mic. Late fusion was performed by learning an additional Naïve Bayes model with the outputs of the individual feature set models as features. It was validated using leave-one-teacher-out cross validation. Overall, we found no added benefit of the single channel fusion models with the exception of the classroom fusion model for Lecture (0.50 vs. 0.54).

Table 1. F_1 score for each target segment for the best performing window size (as subscript)

	Feature Set	Q & A	P & D	Seatwork	Group	Lecture
	Teacher (T)					
1	Timing	**0.61**$_{180}$	0.48$_{180}$	0.41$_{150}$	0.48$_{180}$	0.49$_{150}$
2	NLP	0.51$_{180}$	0.45$_{150}$	0.43$_{120}$	0.44$_{180}$	0.49$_{180}$
3	Channel	0.51$_{150}$	**0.49**$_{120}$	0.46$_{150}$	0.40$_{90}$	0.31$_{60}$
4	Utterance	0.60$_{180}$	0.48$_{150}$	0.32$_{150}$	0.45$_{150}$	0.39$_{30}$
5	T. All	0.56$_{180}$	**0.49**$_{180}$	0.46$_{150}$	0.44$_{150}$	0.46$_{150}$
	Classroom (C)					
6	Timing	0.60$_{180}$	0.48$_{180}$	0.41$_{150}$	0.49$_{180}$	0.49$_{150}$
7	Channel	0.54$_{180}$	0.42$_{30}$	**0.55**$_{150}$	**0.55**$_{150}$	0.50$_{180}$
8	Utterance	0.48$_{90}$	0.37$_{150}$	0.38$_{150}$	0.43$_{30}$	0.15$_{30}$
9	C. All	0.53$_{180}$	0.44$_{30}$	0.49$_{150}$	0.52$_{90}$	**0.54**$_{150}$
10	**T + C**	0.60$_{180}$	0.48$_{180}$	0.49$_{150}$	0.51$_{150}$	0.49$_{150}$
11	Chance	0.21$_{180}$	0.11$_{180}$	0.09$_{150}$	0.22$_{150}$	0.14$_{150}$

Note. *Bold indicates the best approach for each segment. Q&A = Question & Answer, P&D = Procedures & Directions, Seatwork = Supervised Seatwork, Group = Small Group work.*

6.2 Benefits of Combining Both Channels

Next, we combined the seven features together to build a multi-sensor fusion model combining the teacher and classroom channels (Table 1, row 10). This model was built in the same manner as the single channel fusion models described above. We found no added benefit of the multi-sensor (teacher + classroom) fusion model over any of the individual feature sets or the single channel fusion models. If anything, combining sensors actually degraded classification accuracy.

6.3 Comparison to Chance Baseline

We compared our results to a chance-model that assigned the segment label at the rate of the target segment in the dataset but did so randomly. We repeated this process and averaged the values of precision and recall over 1000 iterations. From these averages, we calculated the chance F_1 score for the target segment. We used the best performing window size from Table 1 for each segment. We note that our models clearly outperformed the chance baseline for all five segments.

6.4 Comparison to Previous Work

In previous work [9], we presented the current state of the art using only the teacher's mic. The results of this approach, applied to the current dataset of 56 class sessions for which we have recordings of both the teacher and classroom, are shown in rows 1-3 of Table 1. To examine any potential added benefit of the second classroom mic, we compared the best feature set from the classroom mic (rows 6-9) to the best feature set of the teacher's mic (rows 1-3), and display the percent improvement for each segment in Figure 4. For example, the use of the best classroom feature set (0.55) for Seatwork yielded a 20% improvement over the best teacher feature set (0.46).

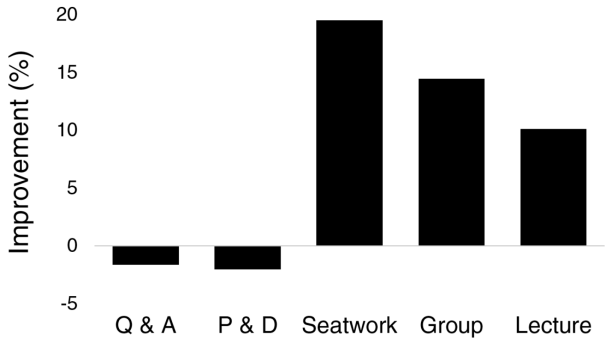

Figure 4. Percent improvement of the best feature set of the classroom model compared to the best feature set of the teacher model for the best performing window size

For Question & Answers and Procedures & Directions, the best performing features sets came from the teacher's mic, and therefore we observed small (< 2%) decreases in performance using either the classroom mic or the multi-sensor fusion model. However, for Seatwork, Group work, and Lecture we found that using a feature set derived from the classroom mic, led to a 20%, 14%, and 10% improvement, respectively, over our prior work, which considered only the teacher mic. Averaged across the five segments, the use of the classroom mic led to an 8% improvement over our prior work, a result that is primarily attributable to improvements in the identification of Seatwork, Group work, and

Lecture. Taken together the present findings validate the need for the additional classroom microphone for the identification of student-centric instructional segments of Seatwork and Group work.

7. DISCUSSION

We considered the task of automatically identifying instructional segments from live classrooms using audio recordings from two rather different mics. This is a challenging task as we are drawing from two uninterrupted channels of audio in order to make high-level predications on instructional activities at specific moments during the class session. Furthermore, our recording from the classroom mic is often noisy and does not always capture the speech of individual students with sufficient fidelity. We consider our results to be significant given the fact that the instructional content discussed in classrooms represents high-level multiparty discourse, but our system did not have the benefit of an accurate text transcript, recordings of individual students, nor video recordings. Instead, it used only low-level features derived from audio recordings of the classroom and teacher.

The present multi-sensor approach that combined a high-fidelity teacher mic with a low-fidelity classroom mic achieved improvements over a previous approach that relied on the teacher mic alone. Overall, we found that using features from the classroom mic led to a 8% improvement across all segments over the state of the art approach [7] that considered teacher mic alone. In fact, the inclusion of the classroom mic led to improvements for three segments and negligible reduction in performance for other two. Specifically, these results demonstrate the utility the classroom audio stream for detection of the student-centric segments of Seatwork, Group work, and Lecture compared to the use of the teacher mic alone. Models using the classroom features also outperformed the multi-sensor fusion model that led to no improvement on this task. This result implies that we could consider using only the classroom mic for the present task of instructional segment classification. However, a multi-sensor approach that utilizes the high-fidelity teacher mic is likely needed for additional classification tasks, such as question detection and question property classification [6].

7.1 Summary of Contributions

We note the following three contributions. First, we described a non-invasive method of recording the teacher using a low-cost and portable microphone that does not interfere with the teacher's routine. This is supplemented by a recording of the classroom which captures both the teacher and general classroom activity without identifying individual students. By prioritizing a relatively simple and affordable recording setup, we will more easily be able to facilitate practical deployment in classrooms.

Second, we built and evaluated our system on a large and diverse dataset that covers multiple teachers, schools, and course subjects. We considered all class recordings, despite the potential absence of certain instructional segments in several class sessions. We also addressed complications from undesirable classroom noise. Importantly, we built and validated our models in a teacher-independent manner which increases confidence that our approach generalizes to new teachers and class sessions. We have found our results scale across the set of eight teachers with no indication that our approach overfits to specific teachers.

Third, we studied the influence of seven different feature types for the detection of instructional segments derived from two parallel but different recordings of the classroom environment. Although previous studies have focused on the students [9] or the teacher [3, 7], this is the first study to consider recordings of both the classroom and the teacher for this task.

7.2 Limitations and Future Work

Our study is not without limitations. One limitation is that our data was collected from within a single U.S. state and does not capture larger geographic differences, such as regional accents and dialects [11] or state-wide curriculum requirements that guide the teacher's lesson plans. We note that the differences between curricula across different states and countries may affect the distribution of certain instructional segments, a potential issue we will consider in the future. Also, we have only tested our system in English language classrooms.

We considered only a single classification model (Naïve Bayes) for all segments to facilitate comparison of results across previous experiments. In further experiments, we will explore choosing a different model to classify each of the five segments, as different classifiers likely have different strengths and weaknesses depending on the instructional segment at hand. Furthermore, based on our observations that certain features are more useful in identifying certain instructional segments, we examined only binary models for each segment. As we continue to refine our approach and improve our results, we will explore combining these into a multi-class approach.

As future work, we will explore the use of temporal models, which will enable the inclusion of additional contextual information when making predictions, a potentially important benefit for the present task. In preliminary experiments, we tested Conditional Random Fields, but found reduced accuracy compared to our current approach. We will continue to explore other temporal models, such as bidirectional Long Short-Term Memory Neural Networks and Hidden Markov Models.

In this work, we compared the performance of seven different features sets, combinations of features for teacher, classroom, and multi-sensor late fusion models. While this approach allowed us to compare the utility of individual feature sets, the ideal set of features for any particular segment classifier will likely derive from a subset drawn from multiple individual feature sets. In future work, we will examine empirical feature selection and additional feature engineering to improve accuracy.

7.3 Applications of the Models

The ability to identify instructional segments used by teachers is necessary in order to generate personalized formative feedback for the teacher about their use of class time. Our approach would permit automating such analysis, enabling a cost-effective scalable deployment that would be accessible to many schools and teachers. Using our system, teachers could record their class and receive feedback following the session. Such feedback will afford teachers reflection on their teaching style and better enable collaboration with professional development personnel towards improvement of their pedagogy, ultimately leading to increased student engagement and achievement. It will also facilitate research into effective pedagogy by providing educational researchers with an automated approach to collect and code classroom discourse.

8. ACKNOWLEDGMENTS

This research was supported by the Institute of Education Sciences (IES) (R305A130030). Any opinions, findings and conclusions, or

recommendations expressed in this paper are those of the author and do not represent the views of the IES.

9. REFERENCES

[1] Alibali, M.W., Nathan, M.J., Wolfgram, M.S., Church, R.B., Jacobs, S.A., Johnson Martinez, C. and Knuth, E.J. 2014. How teachers link ideas in mathematics instruction using speech and gesture: A corpus analysis. *Cognition and Instruction*. 32, 1 (2014), 65–100.

[2] Applebee, A.N., Langer, J.A., Nystrand, M. and Gamoran, A. 2003. Discussion-based approaches to developing understanding: Classroom instruction and student performance in middle and high school English. *American Educational Research Journal*. 40, 3 (2003), 685–730.

[3] Blanchard, N., D'Mello, S., Nystrand, M. and Olney, A.M. 2015. Automatic classification of question & answer discourse segments from teacher's speech in classrooms. *Proceedings of the 8th International Conference on Educational Data Mining (EDM 2015), International Educational Data Mining Society* (2015).

[4] Caughlan, S., Juzwik, M.M., Borsheim-Black, C., Kelly, S. and Fine, J.G. 2013. English teacher candidates developing dialogically organized instructional practices. *Research in the Teaching of English*. 47, 3 (2013), 212.

[5] Chawla, N.V., Bowyer, K.W., Hall, L.O. and Kegelmeyer, W.P. 2002. SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*. (2002), 321–357.

[6] D'Mello, S.K., Olney, A.M., Blanchard, N., Samei, B., Sun, X., Ward, B. and Kelly, S. 2015. Multimodal capture of teacher-student interactions for automated dialogic analysis in live classrooms. *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction* (2015), 557–566.

[7] Donnelly, P.J., Blanchard, N., Samei, B., Olney, A.M., Sun, X., Ward, B., Kelly, S., Nystrand, M. and D'Mello, S.K. 2016. Automatic teacher modeling from live classroom audio. *Proceedings of the 24th Conference on User Modeling, Adaptation and Personalization (UMAP 2016)*. ACM.

[8] Drugman, T. and Stylianou, Y. 2014. Maximum voiced frequency estimation: Exploiting amplitude and phase spectra. *Signal Processing Letters, IEEE*. 21, 10 (2014), 1230–1234.

[9] Gates Foundation 2013. *Ensuring fair and reliable measures of effective teaching: Culminating findings from the MET project's three-year study—Policy and practitioner brief*. Bill & Melinda Gates Foundation Seattle, WA.

[10] Goldman, R., Pea, R., Barron, B. and Derry, S.J. 2014. *Video research in the learning sciences*. Routledge.

[11] Hall, J.K. 2008. Language education and culture. *Encyclopedia of language and education*. Springer. 45–55.

[12] Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P. and Witten, I.H. 2009. The WEKA data mining software: an update. *ACM SIGKDD explorations newsletter*. 11, 1 (2009), 10–18.

[13] Juzwik, M.M., Borsheim-Black, C., Caughlan, S. and Heintz, A. 2013. *Inspiring dialogue: Talking to learn in the English classroom*. Teachers College Press.

[14] Kelly, S. 2007. Classroom discourse and the distribution of student engagement. *Social Psychology of Education*. 10, 3 (2007), 331–352.

[15] Lai, M.K. and McNaughton, S. 2013. Analysis and discussion of classroom and achievement data to raise student achievement. *Data-based decision making in education*. Springer. 23–47.

[16] Lartillot, O., Toiviainen, P. and Eerola, T. 2008. A matlab toolbox for music information retrieval. *Data analysis, machine learning and applications*. Springer. 261–268.

[17] Lewis, D.D. 1998. Naive (Bayes) at forty: The independence assumption in information retrieval. *Machine learning: ECML-98*. Springer. 4–15.

[18] Nystrand, M. 2004. CLASS 4.0 user's manual. *The National Research Center on*. (2004).

[19] Nystrand, M., Wu, L.L., Gamoran, A., Zeiser, S. and Long, D.A. 2003. Questions in time: Investigating the structure and dynamics of unfolding classroom discourse. *Discourse processes*. 35, 2 (2003), 135–198.

[20] Olney, A., Louwerse, M., Matthews, E., Marineau, J., Hite-Mitchell, H. and Graesser, A. 2003. Utterance classification in AutoTutor. *Proceedings of the HLT-NAACL 03 workshop on Building educational applications using natural language processing-Volume 2* (2003), 1–8.

[21] Oviatt, S. and Cohen, P.R. 2015. The paradigm shift to multimodality in contemporary computer interfaces. *Synthesis Lectures On Human-Centered Informatics*. 8, 3 (2015), 1–243.

[22] Prieto, L.P., Sharma, K., Dillenbourg, P. and Jesús, M. 2016. Teaching analytics: Towards automatic extraction of orchestration graphs using wearable sensors. *Proceedings of the Sixth International Conference on Learning Analytics & Knowledge* (2016), 148–157.

[23] Resnick, L.B. 2010. Nested learning systems for the thinking curriculum. *Educational Researcher*. 39, 3 (2010), 183–197.

[24] Samei, B., Olney, A., Kelly, S., Nystrand, M., D'Mello, S., Blanchard, N., Sun, X., Glaus, M. and Graesser, A. 2014. Domain independent assessment of dialogic properties of classroom discourse. *Proceedings of the 7th International Conference on Educational Data Mining (EDM 2014) International Educational Data Mining Society* (2014).

[25] Samei, B., Olney, A.M., Kelly, S., Nystrand, M., D'Mello, S., Blanchard, N. and Graesser, A. 2015. Modeling classroom discourse: Do models that predict dialogic instruction properties generalize across populations? *Proceedings of the 8th International Conference on Educational Data Mining (EDM 2015), International Educational Data Mining Society*. (2015).

[26] Sottilare, R.A., Graesser, A., Hu, X. and Holden, H. 2013. *Design Recommendations for Intelligent Tutoring Systems: Volume 1-Learner Modeling*. US Army Research Laboratory.

[27] Wang, Z., Miller, K. and Cortina, K. 2013. Using the LENA in teacher training: Promoting student involvement through automated feedback. *Unterrichtswissenschaft*. 4, (2013), 290–305.

[28] Wang, Z., Pan, X., Miller, K.F. and Cortina, K.S. 2014. Automatic classification of activities in classroom discourse. *Computers & Education*. 78, (2014), 115–123.