Automatic Teacher Modeling from Live Classroom Audio

Patrick Donnelly[1], Nathan 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; [2]University of Memphis
[3]University of Wisconsin, Madison; [4]University of Pittsburgh
118 Haggar Hall, Notre Dame, IN, 46556, USA
pdonnel4@nd.edu

ABSTRACT
We investigate automatic analysis of teachers' instructional strategies from audio recordings collected in live classrooms. We collected a data set of teacher audio and human-coded instructional activities (e.g., lecture, question and answer, group work) in 76 middle school literature, language arts, and civics classes from eleven teachers across six schools. We automatically segment teacher audio to analyze speech vs. rest patterns, generate automatic transcripts of the teachers' speech to extract natural language features, and compute low-level acoustic features. We train supervised machine learning models to identify occurrences of five key instructional segments (Question & Answer, Procedures and Directions, Supervised Seatwork, Small Group Work, and Lecture) that collectively comprise 76% of the data. Models are validated independently of teacher in order to increase generalizability to new teachers from the same sample. We were able to identify the five instructional segments above chance levels with F_1 scores ranging from 0.64 to 0.78. We discuss key findings in the context of teacher modeling for formative assessment and professional development.

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

1. INTRODUCTION
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 student thought and analysis. The dialogic approach to classroom instruction positively correlates with student engagement [16] and achievement [2, 24]. For example, in a two year, large-scale study of dialogic instruction, Nystrand et al. coded classroom activities for 256 class sessions, covering 2,141 students across 25 schools [23]. After controlling for gender, race/ethnicity, socioeconomic status, school type (e.g., urban/rural, public/private), grade level, and prior achievement, a dialogic-oriented instructional style had positive effects on achievement. In particular, the proportion of time spent on discussion, open-ended questions with no scripted response, and instances of uptake (e.g., follow-up questions) correlated with student achievement [16, 22, 24]. These findings were replicated by another large-scale study of 974 students from 19 different schools across five states [2].

Despite these pedagogical benefits of dialogic instruction, classroom instruction continues to be dominated by traditional teacher-centric instructional techniques such as lecture, recitation, and seatwork [4]. But it need not be this way. Research has demonstrated that the quality of classroom instruction can be enhanced with teacher training programs [6], suggesting that dialogic instruction can be formatively assessed by classroom observations and improved via teacher professional development programs [15]. For example, research has demonstrated that discussing data-driven analysis of classroom practices with teachers correlates with student achievement [17].

The ability to provide teachers with qualitative and formative feedback on their instruction is paramount to improving and refining their teaching strategies over time. Regrettably, current efforts to assess the quality of classroom discourse rely on manual coding by trained observers, a labor and cost intensive endeavor that cannot be deployed practically, broadly, nor uniformly.

To address this critical bottleneck, this study is part of a large multi-disciplinary project that analyzes classroom instructional practices towards the goal of automatic analysis of classroom discourse. The automation of such analysis would lead to the development of a *teacher model*, for use in personalized assessment and professional development. In line with this, we present an approach to automatically identify key instructional segments (e.g., Question & Answer or Lecture) in live classrooms based solely on audio of teachers' speech.

1.1 Related Work
The automatic analysis of text and discourse is a frequently studied research problem in education, such as in automatic essay analysis [12], evaluation of online discussions [20], plagiarism detection [3], or dialog-based intelligent tutoring systems [26]. The focus, however, has been on the student not on the teacher. There is a long research history on the use of audio (and video) to study instructional practices and student behaviors in live classrooms [1, 11] - most notably see [10]. However, the recorded signals are typically processed by humans; automatic analyses of classroom video and audio are few and far between. Thus, while there is an active field of automatic student modeling (or learner modeling) [29], the complementary field of teacher modeling is just beginning to emerge.

The initial attempt at the automatic identification of components of instructional discourse from audio recordings appeared in 2013 by Wang et al. [30, 31]. The authors adapted the Language ENvironment Analysis (LENA) system [9], an expensive

UMAP '16, July 13-17, 2016, Halifax, NS, Canada
© 2016 ACM. ISBN 978-1-4503-4370-1/16/07…$15.00
DOI: http://dx.doi.org/10.1145/2930238.2930250

proprietary microphone intended to be worn by preschool age children, to analyze teacher instruction. They recorded 608 hours of classroom audio from 12 teachers in 1st to 4th grade mathematics classes. They divided the recorded audio into 30 second segments. Two trained coders listened to each segment and annotated the dominant classroom activity: teacher lecture, class discussion, or student group work. They also provided a level of confidence for their annotations. Working independently, the coders achieved an agreement level of 83% (Cohen's kappa, $\kappa = 0.72$). The authors 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 authors trained their classification model using the segments with highest confidence of one coder (62% of the data) and tested on the annotations on all segments of the second coder. Therefore, the same audio segments appeared in both the training and testing sets, albeit using different but highly correlated (83% similar) annotations. Second, the authors did not validate their model independently of the teacher, permitting examples from each teacher to appear in both the training and test sets. Therefore, it is difficult to ascertain if their model was successful in identifying components of the classroom activity or merely adjusting to patterns of speech of specific teachers. Third, all coding was completed offline solely based on the audio recording, thereby losing important visual contextual clues that would be available during a live-coding session. Finally, the authors consider 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).

1.2 Current Contributions

In this paper we describe a low-cost, non-invasive approach to analyze teacher instructional activities in live class sessions. As a proof-of-concept, we previously explored the automatic detection of Question & Answer segments on a dataset of 21 class sessions obtained from three teachers [5]. Using only recordings of teacher audio, we extracted 11 features pertaining to the timing of speech and rest patterns. We achieved an overall accuracy of 67% (AU-ROC of 0.78) validated in a teacher-independent fashion.

Since the eventual goal of this research is to enable wide-scale deployment across many teachers and schools, we design our system with the following design criteria: practicality, generalizability, and scalability [7]. In terms of practicality, the system must be usable by researchers and teachers with minimal training, and non-invasive as to not interfere with the teacher's instruction nor distract students in any way. For this reason, we focus on recording the teachers, unlike approaches that record individual students [9]. It must be economically affordable so that a typical school can afford the system. Additionally, it must not have a human labor cost in that it should run autonomously without human monitoring. Second, the system must be able to generalize to new teachers, classrooms, and domains, hence, we must avoid heavily tuning to the specific teachers or classrooms when training the models. Finally, the system should be flexible enough to operate in a variety of classroom setups and should scale to larger classrooms with minimum loss in fidelity.

The present study advances previous work (see above) and this proof-of-concept in several novel ways. First, we collected the largest dataset in this domain to date, drawn from recordings of multiple teachers across different schools coupled with annotations coded live during each class session. We explore features not previously used in this domain, including analysis of automatic speech transcriptions and acoustic features. We then train supervised classification models to identify five different key instructional segments based on audio recordings of the teacher, validated independently of the teacher and intended to generalize to new teachers.

2. DATA COLLECTION

Data was collected from U.S. middle school literature, language arts, and civics classes. Over the course of three semesters, data was collected from 76 class sessions, covering eleven different teachers (three male, eight female) across six schools. The teachers were not coached in the practice of dialogic instruction and were asked to carry out their normal lesson plan, allowing the capture of an unbiased real-world sample of teachers' instructional practices.

Each teacher wore a wireless microphone to capture their speech. Based on previous work [7], 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.

Each class session lasted between 30 minutes to 90 minutes, depending on the school, with an average class length of 60:25 minutes. These recordings, totaling over 76 hours, capture the gamut of events typical in a classroom, from focused instruction to distracting interruptions.

2.1 Coding Classroom Discourse

The Nystrand and Gamoran classroom coding scheme [21] considers a hierarchy of classroom events, ranging from general to more specific: (1) episodes refer to the general topic being addressed in the class (e.g., "*the Civil War*"); (2) instructional segments represent one the 17 possible instructional activities used to implement the episode (e.g., Lecture, Discussion), and (3) individual questions asked by teachers or students during some instructional segments. We focus on the second level of this coding scheme, the automatic identification of instructional segments.

An observer who was trained in the Nystrand and Gamoran scheme was present during each recorded class session. The observer used software specifically developed for live coding of classroom discourse to mark episodes, instructional segments, and teacher's dialogic questions as they occurred. Live coding allowed the observer to utilize visual information, which ostensibly yields additional information to contextualize the coding. For example, the coder may observe that students are working on a task in small groups rather than individually, an assessment that may be difficult to determine from the audio recording alone.

There were three trained observers in this study. Each class session was coded by one observer, whose coding was subsequently verified by a second trained observer at a later time. Disagreements were discussed and the coding refined until both observers reached complete agreement. The instructional segments noted by the coders form the ground truth used to evaluate our classification models.

2.2 Analysis of Instructional Segments

We focus on detecting the five most frequent segments that individually comprised at least 10% of the data: Question & Answer (21%), Procedures and Directions (20%), Supervised Seatwork (12%), Small Group Work (11%), and Lecture (11%). Ironically, Discussion, an instructional segment important to student success,

represents 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 22%.

There are eleven additional types of instructional segments that occur less frequently, such as an occasional distraction, the discipline of a student, a test or quiz, or students engaging in silent reading. Individually, these segments are rare, but together they comprise 24% of the dataset. Although we do not build models for these segments, we retain them in our dataset as a Miscellaneous category.

We refer to [21] 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. These segments may feature pre-scripted test questions by the teacher or they may be open-ended, providing the opportunity to transition into a more in-depth Discussion segment. 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, occasionally supplemented with video.

Error! Reference source not found. shows the class time spent on each of the five key instructional segments by each teacher. The individual teachers divide time differently, a reflection of their unique style. For example, two teachers did not assign any Small Group Work and three teachers did not spend any lesson time on Supervised Seatwork. There was also considerable variation over the different class sessions, even within each teacher, which reflects differences in lesson plans each day. In general, an individual teacher or class session may not contain all five instructional segments, a challenge we discuss in our results.

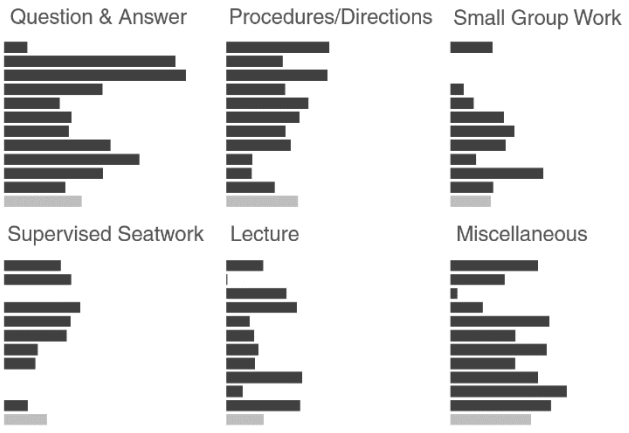

Figure 1. Proportion of time spent by each of the eleven teachers on each of the instructional segments. For each teacher (each row), the total time across the six segments sums to 1.0. The proportion of total time spent on the segment within the dataset is shown last in gray.

3. MODEL BUILDING

In this section, we discuss our approach to building classification models for the identification of the key instructional segments. We present our approach to segmenting the teacher's audio channel and generating automatic speech recognition transcripts, followed by discussion of our approach to partitioning the classroom recordings into windowed instances for classification. Finally, we present our feature extraction scheme and present our classification models.

3.1 Partitioning Audio into Windows

A system that is to be deployed in classrooms will not have the benefit of human coders present and we will be unable to determine the boundaries of the instructional segments. Although we could potentially build detectors to automatically infer segment boundaries, this itself is an unsolved research problem beyond the scope of this study. Therefore, we divided the recording of each class session into consecutive non-overlapping windows of time for classification. We examined non-overlapping windows to consider each moment of class time only within a single windowed instance 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 of the classroom activity 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 spans 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 2. This approach, although an imperfect generalization, allows tracking the broader picture of the teacher's time. In particular, the average segment is 2.9 minutes long, which we use to inform our selection of possible window sizes. Specifically, we explored windows sizes ranging from 30 seconds to 5 minutes.

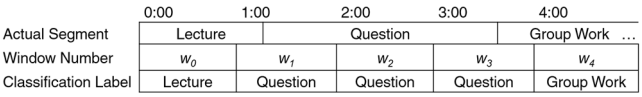

Figure 2. Example of the windowing scheme for a sample of five minutes of class time considering a 60 second window.

3.2 Utterance Segmentation

Each recording represents an uninterrupted channel of teacher audio lasting the duration of the class session. In order to analyze teachers' instructional practices, we must subdivide the audio signal into smaller audio chunks, each of which ideally represents an utterance spoken by the teacher. We adopted a method developed in [7] for teacher utterance extraction.

Patterns of speech and rest differ between teachers as do unintentional noises such as breathing or coughing. Therefore, we employed a general method to segment utterances to avoid overfitting to specific teachers, potentially increasing the ability to generalize to new teachers. 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, which was empirically tuned in previous work [5]. Whenever the amplitude remained below the threshold for at least one second, we identified this as a moment of silence in the recording and used this silence as a breakpoint from which to partition the recording into the smaller utterances (called *potential utterances*). This approach

is not without limitations, as each utterance may contain multiple ideas, or a single idea may be spread across several utterances.

Next, we analyzed this set of potential utterances in order to retain those that contain the teachers' speech and discard others (e.g., background noise, heavy breathing). To identify the utterances containing speech, we passed each through the Microsoft Bing automatic speech recognition (ASR) system [7]. If the ASR identified any speech within the potential utterance, we retained it as a speech utterance, otherwise we discarded it. In a validation study using 1000 randomly-sampled potential utterances, we achieved high levels of both precision (96.3%) and recall (98.6%) using this method. This resulted in an F_1 score of 97.4% [7], which we deemed sufficiently accurate for the purposes of this study.

Using this process on our dataset of 76 classroom recordings yielded 40,138 candidate utterances, 23,610 (59%) of which were retained as containing speech. The average length of these speech utterances is 5.24 seconds (SD = 8.17), however 2% of the utterances last over thirty seconds in length; for example, when the teacher makes a long statement without pausing.

3.3 Feature Extraction

We extracted features from each of the windows (see Figure 2 above) to create the instances used to train and test our classification models. We explored three different types of features, two of which have not yet been explored for this task. Specifically, timing features capturing patterns of the teacher speaking and pausing have been previously considered [5, 31]. We complimented these with novel features generated from natural language processing of speech transcriptions of the teacher, no small task due to the noisy nature of the classroom. These features allow consideration of the specific words a teacher speaks, beyond the mere timing of speech. We also added acoustic features generated from the recording of the teacher. We chose these features to potentially help differentiate between speaking and silence, for example, or between the difference in classroom noise generated by a single speaker and the louder moments when many voices speak simultaneously, such as occurs during Small Group Work.

Utterances Timing Features: We analyzed the timing of the extracted teacher utterances described in Section 3.2. For each partitioned window of time, we identified any speech utterances present within the windows. 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 any time between utterances as rest, we constructed a sequence of speech and rest. We then extracted six features from 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. In a similar manner, we extracted the same six features from the timing of the rest patterns. We added in one more feature representing the normalized temporal position of the window proportionate to the total length of the classroom recording, resulting in a total of 13 features.

Natural Language Features: In prior work, we evaluated several different ASR engines on data recorded in the noisy classroom environment [4, 7]. We considered two metrics: word error rate (WER), a word level edit distance comparing the ASR and human transcripts, and simple word overlap (SWO), a measurement of proportion of words found in both transcripts. Bing ASR achieved a WER of 0.52 and a SWO score of 0.62. Although outperformed by the Google ASR engine, we selected Bing given its ability to freely transcribe large volumes of audio, an important consideration for broader deployment.

Given an ASR transcript generated for each teacher speech utterance, we must extract meaningful language features that ostensibly capture differences in instructional segment. For this task, we employed a natural language feature tagger [25] that was specifically designed to classify questions and has been validated in studies of classroom discourse [27, 28]. 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 since the course material is not likely to repeat between teachers. We extracted 37 natural language features, including counts of parts of speech (e.g., adjectives, nouns) and counts of particular words (e.g., *what, how, why*). 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 include both the sum and mean of the 37 natural language features to attempt to capture differences of use within the window, for a total of 74 features. Although potentially correlated, the sum of the feature counts the number of times the feature occurred within the time window whereas the mean tracks the use of the feature averaged by number of utterances in the windows. For example, consider the question word "*why*." The word may appear multiple times in a single long utterance, such as during a Lecture, or it may appear across of sequence of speech and rest by teacher, potentially signally a Question & Answer segment in which the teacher's speech alternates with student responses.

Acoustic Features*:* Lastly, we extracted a set of features based on the acoustic properties of the audio signal using the Music Information Retrieval toolbox for Matlab [18]. Unlike the other aforementioned modalities, these features were not extracted from segmented teacher utterances but directly from the window of audio. We did not calibrate the features by individual teacher to encourage generalization to new teachers in the future. These features include common descriptors that characterize volume, spectra, and the frequency curve of the signal. We include the following features: seven 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 is 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 representation of the short-term power spectrum. Additionally, we included 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 time window.

3.4 Supervised Classification

We generated 117 features in total for each windowed partition of the audio recording. These features were used to train supervised classification models to identify instructional segments. As noted in Section 2.2, there was considerable data imbalance due to an infrequent occurrence of some segments and the high variance in use between different teachers and 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 or one of the other potential segments, whether another common segment or one of the 11 infrequently occurring Miscellaneous segments.

We considered the Naïve Bayes classifier using the WEKA machine learning toolbox [14]. 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) and because of its popular and successful use as a baseline classifier in many domains [19].

4. EXPERIMENTS AND RESULTS

All experiments were conducted with a leave-one-teacher-out cross validation. For each of the 11 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 ten teachers. This process was repeated for each teacher such that each teacher appeared in the test set only once, and the results were calculated as the average of the 11 folds. This approach allows better generalization to new teachers, preventing the classification models from overfitting based on characteristics of individual teachers.

In terms of metrics, accuracy, or recognition rate, is not an ideal measure when base rates between class labels are highly skewed, as they are in our data. Therefore, we evaluated the efficacy of our binary target segment vs. all others classifier models by examining the F_1 score, a balance of precision and recall, for the class label of interest (i.e. the target segment such as Question & Answer). This ensures that we focus on the model's ability to detect the segments of interest, which was always the minority label, rather than prioritizing the dominant class label (i.e. the other category).

4.1 Comparing Window Time

The size of an analysis window is an important design choice as it determines the temporal resolution of our predictions. While a shorter window will yield more fine-grained predictions, a longer window allows the consideration of more information for each prediction. In this experiment, we varied the window size ranging from 30 seconds to 300 seconds in increments of 30 seconds. The F_1 scores for each target segment are shown in Figure 3.

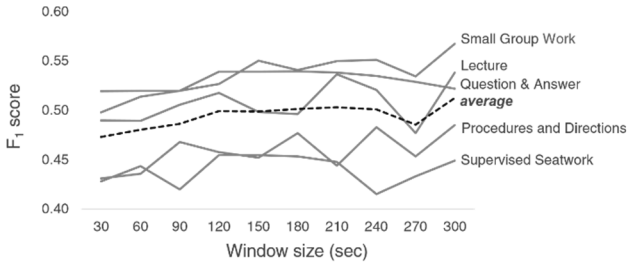

Figure 3. The F_1 score of the target segment label for each window size. The average score of the five classifiers is represented as a dashed line.

The results indicated that all five classifiers had lower performance on short window times, such as 30 seconds. Although performance generally increased with window size, longer windows resulted in fewer instances per class session. Furthermore, given the majority takes-all segment labeling approach described in Section 3.1, longer windows risk masking short occurrences of instructional segments. This effect is undesirable because it loses information about the use of class time.

The results indicated that the optimal classification window might need to be varied between classification models for different segments. In future work, we will explore tuning the window size depending on the type of instructional segment to be classified. For the reminder of the experiments, we focused on a window size of 120 seconds, which showed improved performance for all classifiers over shorter window sizes while not exceeding the average length of segments (176 seconds). A window size of 120 seconds yielded 2,254 instances for classification across the 76 class sessions.

4.2 Comparing Feature Types

In our second experiment, we explore the relative effectiveness of the different feature types described in Section 3.3 in the classification of instructional segments. We trained a separate Naïve Bayes classifier for each of the three feature types (timing, NLP, or acoustic), and a fourth model which fused all three sets of features. The results are shown as Figure 4.

We observe that NLP features were most successful in detecting Procedures and Directions, perhaps unsurprisingly as these segments feature common patterns of imperative instructions provided by the teacher. Timing of teacher's speech and rest patterns aided in the classification of Question & Answer segments. The acoustic features were notably less useful in the classification of Small Group Work and Supervised Seatwork. As the present work only considers a recording of the teacher, it is likely that not enough information is available during these student-focused segments. Furthermore, the acoustic features may have difficulty in generalizing between male and female teachers as our current dataset contains more examples of female teachers.

We compared the average F_1 score across segments by feature type: timing (0.49), NLP (0.53), and acoustic (0.36) to the average score using all features (0.52). Although the average score of NLP features trivially exceeds the score of all features, we observe that the boost only comes from the Procedures and Directions classifier. Therefore, we consider all the features in the remaining experiments, although we will investigate additional feature engineering in future work.

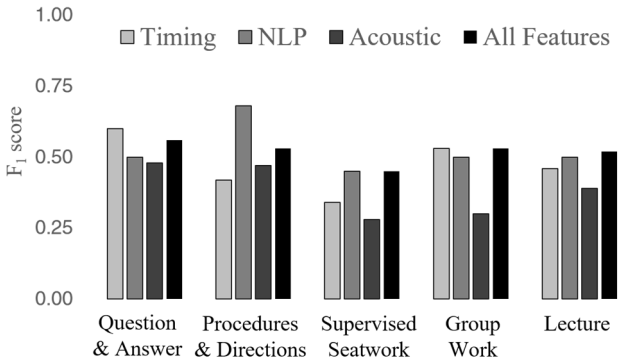

Figure 4. The F_1 score of the target segment label for each of the three feature types and the combination of all features.

4.3 Individual Feature Analysis

Following the analysis of feature category in the previous section, we explored the utility of each of the 117 individual features. In this experiment, we trained a binary classifier for each single feature and for each instructional segment. We ranked the features by the F_1 score for the target segment to explore the contribution of individual features. There was no clear pattern of features that worked best across the different segments. Overall, no single

feature achieved an F_1 score greater than 0.25. This is unsurprising as we did not expect a single feature to dominate over the combination of many features.

For Question & Answer segments, the ten most useful features were all NLP features and included features such as the number of occurrences of certain question words (e.g., "*why*", "*what*") and the number of proper nouns. This is an encouraging result, demonstrating the utility of our NLP features despite the noise introduced by the imperfect ASR transcription. Furthermore, this demonstrates the timing features that capture teacher speech-rest patterns, used in previous work [5, 31], should be supplemented with additional feature modalities. Procedures and Directions also benefited most from NLP features, although these features differed from those useful to detect Question & Answer. For Supervised Seatwork, we note that acoustic features account for seven of the top ten features. This is an encouraging result and support our hypothesis that student-focused segments may be difficult to identify based on the teacher's speech, as the teacher may be silent for extended periods of time. No single feature alone achieved success in identifying Small Group Work from the other segments. However, since we are able to classify Small Group Work above chance levels using all features, this result underscores the utility of combinations of different features. The Lecture classifier best benefited from timing features, particularly those that described the length of moments of rest, which may necessary to determine non-Lecture segments in which the teacher is silent for extended periods of time.

4.4 Final Classification Results

Informed by the previous experiments, we trained a set of Naïve Bayes classifiers, considering a window of 120 seconds and using all 117 features to generate 2,254 windowed instances for classification. In Table 1, we report the overall F_1 score to provide a measure of the binary classifiers' performance across both the class labels. The target F_1 score tracks the performance of the classifier only on the target label of interest (e.g., Lecture). We include recognition rate (accuracy) and AU-ROC for comparison to other studies.

Table 1. Classification results for each of the five instructional segments. Target F_1 refers to the F_1 score of the segment of interest, listed in each row, while overall F_1 represents the weighted score considering both labels.

	Target F_1	Overall F_1	Rec. Rate	AU-ROC	*Target chance*
Question & Answer	0.55	0.64	0.60	0.76	*0.31*
Procedures/ Directions	0.47	0.64	0.60	0.72	*0.27*
Supervised Seatwork	0.45	0.67	0.59	0.65	*0.19*
Group Work	0.53	0.65	0.56	0.70	*0.19*
Lecture	0.52	0.78	0.71	0.58	*0.16*
Average	0.50	0.68	0.61	0.68	*0.23*

For a binary dataset that contains an equal distribution of labels, an F_1 score of 0.50 represents chance, and reflects a random assignment of the two labels. However, for datasets containing a large imbalance in the dataset, such as ours, the level of chance prediction is not as straightforward. For comparison to our target F_1 score, we calculated chance levels as follows. We considered chance-level precision as the precision of a classifier that always selects the target segment, which yields a precision that matches that proportion of the target segment in the dataset (see Section 2.2). We considered chance-level recall as prediction rate of the target label (e.g., Lecture) for each segment classifier. Using these values, we calculated an F_1 baseline for chance prediction of the target segment, shown in Table 1. We define chance in this manner to emulate classification with the same prediction rate as our models on a dataset reflecting the same distribution as our data.

The target F_1 score reveals the efficacy of predicting the minority class labels, which correspond to each of our key instructional segments. For all segments, we were able to predict at levels well above our target F_1 chance baseline. However, we were more successful predicting Question & Answer, Small Group Work, or Lecture compared to Procedures and Directions or Supervised Seatwork.

Table 2 presents the confusion matrix for each binary classifier as a proportion of the total instances. We most readily correctly identified true cases of Question & Answer segments (72% of the time), compared to the target class of the other classifiers. We note that all five classifiers were able to identify their respective segment at levels well above chance, but do suffer from misclassifications. In particular, the classifiers had high false positive rates, in which, for example, a non-Seatwork segment was identified as a Supervised Seatwork segment. This too is likely a consequence of the frequency of occurrence of certain segment types for certain teachers.

Table 2. Confusion matrices of each of the five classifiers. The column headers represent the predicted segment, while the row header denotes the actual segment.

	Actual	Predicted	
Question & Answer		*Q&A*	*Other*
	Q&A	0.72	0.28
	Other	0.44	0.56
Procedures and Directions		*Directions*	*Other*
	Directions	0.70	0.30
	Other	0.42	0.58
Supervised Seatwork		*Seatwork*	*Other*
	Seatwork	0.63	0.38
	Other	0.44	0.56
Small Group Work		*Group*	*Other*
	Group	0.67	0.34
	Other	0.45	0.55
Lecture		*Lecture*	*Other*
	Lecture	0.59	0.41
	Other	0.27	0.73

The confusion matrices are generally symmetric, with the exception of the Lecture classifier. Here, it appears that we are more successful at detecting non-Lecture than the Lecture segments themselves. This might be because Lecture segments are much more variable as they pertain to the general subject matter of the day, topics unlikely to be visited in other class sessions.

Furthermore, Lecture segments may contain supplementary video or other aspects that add unique challenges to classify based solely on audio recordings.

4.5 Comparison with Previous Work

A direct comparison with previous work is not possible because our dataset, preprocessing steps, feature extraction, and classifiers differ substantially. Nevertheless, we discuss our work in the context of the previous two studies in this domain. On the surface, our Question & Answer segment results are comparable to [5], which considered only Question & Answer segments from a small set of three teachers, reporting an AU-ROC of 0.78. This study reported results from a logistic regression classifier, a classifier which we also considered in our preliminary experiments but discarded as it had a tendency to overfit to the dominant class label and did not scale well to larger sets of features. Furthermore, in our work, we considered Question & Answer and Discussion segments together, rendering it a harder problem while [5] simply discarded Discussion segments altogether.

Wang et al. reported 84% accuracy across a limited set of three possible instructional segments [31]. As we reviewed in Section 1.1, the authors re-used their training examples in their test set, albeit with testing the label given by the other human coder, achieving an accuracy of 84%. Since two coders had an 83% agreement in their annotations, this resulted in highly correlated training and test sets. In comparison, our approach used separate training and test sets validated independently of the teacher. The differences between our results and [31] underscore the need to validate classification models in a manner independent of the teacher or the class session in order to generalize to new teachers and class sessions.

5. DISCUSSION

We considered the task of automatically identifying instructional segments from live classrooms using only an audio recording of the teacher's speech. This is quite a challenging task as we are drawing from a single uninterrupted channel of classroom audio in order to make high-level predications on instructional activities at specific moments during the class session. Although our classification models are not perfect, we are able to detect five individual instructional segments well above chance levels. Despite the fact that the instructional content discussed in classrooms represents high-level discourse, our system did not have the benefit of an accurate text transcript or recordings of individual students. Instead, it used only low-level features derived solely from teacher audio.

5.1 Contributions

Our system fulfills our design goals of practicality, generalizability, and scalability. 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 regular teaching routine. All data processing tasks, including audio capture, automatic speech recognition, feature extraction, and classification can be performed on a standard personal laptop. By prioritizing a simple and affordable technical setup, we will more easily be able to facilitate practical deployment in classrooms. This is a significant advantage over the approach used in [31] which requires expensive and propriety recording equipment and analysis software.

Second, we evaluated our system on the largest and most diverse dataset thus far considered for this task, covering 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 had to handle the difficulties of undesirable noise, such as the persistent heavy breathing of one teacher or distracting background noise from the classroom, in others. Additionally, we focused on the identification of five key instructional segments, extending beyond previous attempts at automatic classification of instructional activities, which considered only a single activity [5] or a limited set of only three activities [31]. Of the five instructional segments considered in this work, Question & Answer segments are the most important component of dialogic instruction as certain types of questions and in-depth discussion sections correlate to increased student achievement [16, 22, 24]. Therefore, it is encouraging that we can more readily identify these Question & Answer segments, although further refinement is needed to reduce the false positive rate.

Third, we studied the influence of three diverse features types for the detection of instructional segments. We considered features derived from natural language processing of ASR transcriptions and non-verbal acoustic features extracted from noisy classroom audio recordings. It is encouraging that the NLP features were successful in identifying certain segments, despite the fact that they are generated from ASR transcriptions, an imperfect process hindered by mumbled speech or ambient background noise.

Most importantly, we built and validated our models in a teacher-independent manner which increases confidence that our approach generalizes to new teachers, schools, or class sessions. We have found our results scale across the set of eleven teachers with no indication that our approach overfits to specific teachers.

5.2 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 phrasing [13] or state-wide curriculum requirements that guide the teacher's lesson plans. We anticipate that different regional accents may not be a significant issue given the wide-spread use of ASR, but this requires empirical confirmation. We have also only tested our system in English language classrooms. Given the proliferation of ASR for many languages, we anticipate our approach will largely extend to other languages, provided an adaption be made to the natural language processing features to suit other languages. Lastly, 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.

Although this work demonstrates encouraging progress towards the goal of automatic analysis of class instruction, significant refinement is necessary to improve the efficacy of our predictions. In particular, our classifier will likely benefit if given additional data apart from the teacher's speech. Recording individual students is impractical with regards to both cost and privacy concerns. Presently, additional data collection is underway which includes a pressure zone microphone to capture general classroom activity. Although the additional microphone is still subjective to the same challenges as the teacher channel, such as classroom noise or imprecise speech transcriptions, this second channel of audio, coupled with the recording of the teacher, would allow modeling teacher-student interactions, potentially yielding stronger insight to the classroom activity in progress.

We also considered only a single classification model (Naïve Bayes) to facilitate comparison of results across experiments. In further experiments, we will explore the use of different classifiers for each of the five segments, as different classifiers likely have

different strengths and weaknesses depending on the instructional segment at hand. Furthermore, as we continue to refine our approach and improve our results, we will explore combining the binary models into a multi-class approach to classification.

Furthermore, we observed that we are better able to identify an instructional segment if it occurred frequently during the class session. There was, however, a tendency to overpredict when a segment did not occur in a classroom session. To address this limitation, we will explore models that consider the instructional segments in the larger context of the class session. For example, we will attempt to first predict if examples of the segment exist in a particular class session, before classifying the individual windowed segments. Although we validated our approach across teachers, we did not explicitly consider the order of the windowed instances within a class session. As future work, we will explore the use of temporal models, such as a hidden Markov models or conditional random fields, which can incorporate information that occurred earlier in the class session when making predictions. This approach enables the inclusion of additional contextual information when making predictions, a potentially important benefit for the present task.

A major difficulty of our task stems from the imbalance of segments of interest across the entire dataset. In order to work towards the goal of achieving a deployable system, we must overcome the challenge of class label imbalance as it reflects the reality of real-world classes. The collection of additional classroom recording is ongoing and this will provide more examples of the various instructional segments and additional teachers.

5.3 Concluding Remarks

We took steps towards automated teacher modeling by identifying teachers' instructional activities from audio data collected in live classrooms. The teacher model will be used to generate personalized formative feedback, which will afford reflection and improvement of their pedagogy, ultimately leading to increased student engagement and achievement.

6. 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.

7. 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] Barrón-Cedeño, A., Vila, M., Martí, M.A. and Rosso, P. 2013. Plagiarism meets paraphrasing: Insights for the next generation in automatic plagiarism detection. *Computational Linguistics*. 39, 4 (2013), 917–947.

[4] Blanchard, N., Brady, M., Olney, A.M., Glaus, M., Sun, X., Nystrand, M., Samei, B., Kelly, S. and D'Mello, S. 2015. A Study of Automatic Speech Recognition in Noisy Classroom Environments for Automated Dialog Analysis. *Artificial Intelligence in Education*. Springer International Publishing. 23–33.

[5] 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).

[6] 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.

[7] 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.

[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] Ford, M., Baer, C., Xu, D., Yapanel, U. and Gray, S. 2008. *The LENA language environment analysis system*. LENA Foundation Technical Report LTR-03-02.

[10] 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.

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

[12] Graesser, A.C. and McNamara, D.S. 2012. Automated analysis of essays and open-ended verbal responses. *APA handbook of research methods in psychology. Washington, DC: American Psychological Association*. (2012).

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

[14] 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.

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

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

[17] 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.

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

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

[20] Mu, J., Stegmann, K., Mayfield, E., Rosé, C. and Fischer, F. 2012. The ACODEA framework: Developing segmentation and classification schemes for fully automatic analysis of online discussions. *International Journal of Computer-Supported Collaborative Learning*. 7, 2 (2012), 285–305.

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

[22] Nystrand, M. 2006. Research on the role of classroom discourse as it affects reading comprehension. *Research in the Teaching of English*. (2006), 392–412.

[23] Nystrand, M., Gamoran, A., Kachur, R. and Prendergast, C. 1997. *Opening dialogue: Understanding the Dynamics of Language and Learning in the English Classroom. Language and Literacy Series*.

[24] 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.

[25] 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.

[26] Rus, V., D'Mello, S., Hu, X. and Graesser, A. 2013. Recent advances in conversational intelligent tutoring systems. *AI magazine*. 34, 3 (2013), 42–54.

[27] 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).

[28] Samei, B., Olney, A.M., Kelly, S., Nystrand, M., Blanchard, S.D.N. and Graesser, A. Modeling Classroom Discourse: Do Models that Predict Dialogic Instruction Properties Generalize across Populations?

[29] 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.

[30] 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.

[31] 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.