Multimodal Capture of Teacher-Student Interactions for Automated Dialogic Analysis in Live Classrooms

Sidney K. D'Mello[1], Andrew M. Olney[2], Nathan Blanchard[1], Xiaoyi Sun[3], Brooke Ward[3], Borhan Samei[2], and Sean Kelly[4]
[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
sdmello@nd.edu

ABSTRACT
We focus on data collection designs for the automated analysis of teacher-student interactions in live classrooms with the goal of identifying instructional activities (e.g., lecturing, discussion) and assessing the quality of dialogic instruction (e.g., analysis of questions). Our designs were motivated by multiple technical requirements and constraints. Most importantly, teachers could be individually mic'ed but their audio needed to be of excellent quality for automatic speech recognition (ASR) and spoken utterance segmentation. Individual students could not be mic'ed but classroom audio quality only needed to be sufficient to detect student spoken utterances. Visual information could only be recorded if students could not be identified. Design 1 used an omnidirectional laptop microphone to record both teacher and classroom audio and was quickly deemed unsuitable. In Designs 2 and 3, teachers wore a wireless Samson AirLine 77 vocal headset system, which is a unidirectional microphone with a cardioid pickup pattern. In Design 2, classroom audio was recorded with dual first-generation Microsoft Kinects placed at the front corners of the class. Design 3 used a Crown PZM-30D pressure zone microphone mounted on the blackboard to record classroom audio. Designs 2 and 3 were tested by recording audio in 38 live middle school classrooms from six U.S. schools while trained human coders simultaneously performed live coding of classroom discourse. Qualitative and quantitative analyses revealed that Design 3 was suitable for three of our core tasks: (1) ASR on teacher speech (word recognition rate of 66% and word overlap rate of 69% using Google Speech ASR engine); (2) teacher utterance segmentation (F-measure of 97%); and (3) student utterance segmentation (F-measure of 66%). Ideas to incorporate video and skeletal tracking with dual second-generation Kinects to produce Design 4 are discussed.

Categories and Subject Descriptors
I.2.7 [Artificial Intelligence]: Natural Language Processing – *discourse, Speech recognition and synthesis, text analysis.*

General Terms
Algorithms, Experimentation, Human Factors.

ICMI 2015, November 9–13, 2015, Seattle, WA, USA.
 ISBN 978-1-4503-3912-4/15/11…$15.00.
DOI: http://dx.doi.org/10.1145/2818346.2830602

Keywords
Multimodal; classroom discourse; dialogic instruction

1. INTRODUCTION
A particular style of classroom discourse, known as dialogic instruction, has been found to be a powerful predictor of student achievement in middle and high-school English Language Arts and Social Studies classrooms [22-24]. Dialogic instruction focuses on the free exchange of ideas and open-ended discussion between teachers and students with the goals of provoking deeper student thought and analysis and more evenly distributing effort among students. The resultant increase in critical thinking and student engagement [13] are the mechanisms by which dialogic instruction leads to improved achievement [8].

In the first major quantitative study of dialogic instruction, Nystrand and colleagues [22] coded discourse practices involving thousands of eighth- and ninth-grade students in a diverse sample of hundreds of English Language Arts classes. Their coding scheme consisted of three key 'tracks,' of increasingly fine granularity: 1) *episodes*, which refer to the activity/topic being addressed in class (e.g., the civil war); 2) *segments*, which represent possible instructional activities used within the episode; (e.g., lecturing, discussion, group work) and 3) *questions* asked by teachers or students embedded within segments. When controlling for gender, race, ethnicity, socioeconomic status, grade level, and prior achievement in writing and reading, they found that dialogically organized instruction (indicated by the following three items) had positive effects on reading comprehension (including both recall and depth of understanding as well as response to aesthetic aspects of literature):

- *Discussion*: Open exchange of ideas among at least three participants lasting longer than 30 seconds;
- *Authentic teacher questions*: Open-ended questions without a pre-scripted answer rather than known-answer test questions; and
- *Uptake*: Speaker's incorporation of a previous answer into a subsequent question (e.g., follow up question – "what did you mean by that")

These results were replicated by Applebee, et al. [2] in a 1-year study of 974 students in 64 middle and high school English class-rooms in 19 schools across 5 states. This body of research has also demonstrated that the quality of dialogic instruction can also be enhanced with teacher training programs, suggesting that dialogic instruction can be formatively assessed (by classroom observations) and improved via teacher professional development programs.

Unfortunately, the research is difficult to scale and put into practice because dialogic instruction must be manually coded by trained observers. Even with performance support tools developed by Nystrand and colleagues, including live coding CLASS 4.24

software [21], it still requires approximately 4 hours of coding time per 1 hour of classroom observation. This is an unsustainable task for scalable research and for providing day-to-day feedback for teacher professional development.

To address this key limitation, the present project (CLASS 5) is focused on automatically analyzing classroom discourse as a means of providing feedback to researchers, teachers, teacher educators, and professional development personnel. Our approach uses automatic speech recognition (ASR), natural language understanding, and machine learning, which requires collection of large volumes of data in authentic classroom settings. This poses a number of technical and logistical challenges. The purpose of this paper is to describe our data collection designs and corresponding analyses techniques to meet these challenges.

Our initial designs and experiments focus on recording multiple streams of audio in live classrooms. We emphasized audio in this initial stage of work due to the intended application and because audio is an important component modality for multimodal interactions. Once a satisfactory audio recording solution has been developed, we extend the basic designs to incorporate visual information.

In the remainder of this section, we discuss the technical requirements and constraints that motivated our design, briefly consider related work in this area, followed by an overview of our approach.

1.1 Technical Requirements and Constraints

Our design is guided by the following technical requirements and constraints:

1. *Usability*: The system is intended for use by researchers, teachers, teacher educators, and professional development personnel. These individuals might have limited technical proficiency, so no advanced technical skills should be required.
2. *Economic Cost*: Equipment cost should be between ($500 to $1,000) so that the average school can afford the system.
3. *Human Cost*: The system should run autonomously after a brief (<5 minute) setup period. No human monitoring should be necessary during data collection.
4. *Scalability*: The system should be suitable for average public classrooms in the U.S., where class sizes are an average of 21.2 students and 26.8 students for public elementary and secondary schools, respectively [20].
5. *Flexibility*: The design should accommodate common classroom designs including traditional lecture (teacher up front with students seated in rows), small group work (students sitting around tables with teacher walking around), and designs promoting discussion (students sitting around a circle/oval with teacher on the perimeter).
6. *Non-intrusiveness*: The design should be minimally intrusive in that normal classroom instruction should not be interrupted.
7. *Visual information*: Video is desirable but not required. Further, due to privacy concerns, video data must either be processed in real time so the actual images are discarded or transformed into a non-personally-identifiable form.
8. *Audio information*: The teacher can be mic'ed but students cannot be individually mic'ed.
9. *Teacher audio quality*: High quality teacher audio is required. Teacher quality should be sufficiently high for teacher utterance segmentation and ASR.
10. *Student audio quality*: Student audio quality should be sufficient for detection of student speech. ASR on student audio is desired but not required.
11. *Stream synchronization*: Audio (and potentially video) data streams need to be synchronized with millisecond accuracy. These data streams need to be synchronized with the live coding that occurs in the classrooms within a 500 ms margin of error (during data collection for development of CLASS 5).
12. *Identity resolution*: Identity only needs to be resolved to the level of teacher vs. student(s), but not to individual students.

1.2 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, 11] - most notably see [9]. However, the recorded signals are typically processed by humans; automatic analyses of classroom video and audio are few and far between. Thus, the literature did not provide much guidance on how to instrument a live classroom in order to collect data of sufficient quality for automated analysis.

The closest related work is research by Wang, et al. [30], who recorded classroom audio in 1st to 3rd grade math classes with the goal of automatically analyzing instructional segments in order to boost the level of discussion in these classes. Their data collection design consisted of 11 teachers wearing the LENA (Language Environment Analysis [7]) recorder (a small nonintrusive device) while teaching mathematics classes. LENA is a wearable system which records and measures surface-level aspects of language produced by and directed at very young children. The LENA propriety software uses differences in volume and pitch in order to assess when teachers were speaking, students were speaking, speech was overlapping, or there was silence.

Wang, et al. [30] report impressive accuracy rates ranging from 0.95 to 0.99 for teacher utterance segmentation. However, accuracy rates for student speech segmentation were lower due to the system misclassifying student speech as teacher speech. They attribute this to small differences in pitch between teachers and elementary school students since LENA was optimized to differentiate between caregivers and very young children. In particular, LENA was designed to analyze speech produced by children between 2-48 months in age [7], but was applied for a much older age group (1st to 3rd graders) in the Wang study. To address this, they adapted the underlying algorithm, which resulted in improved student speech segmentation accuracy rates ranging from 0.70 to 0.86.

The Wang, et al. [30] study is pioneering in a number of ways. Their approach satisfies the following subset of our technical requirements and constraints: usability, human cost, scalability, non-intrusiveness, audio information, student audio quality, and identity resolution. It is agnostic with respect to visual information and stream synchronization. Adherence to the flexibility requirement is unknown. Their approach is somewhat limited in terms of economic cost because it requires customized commercial hardware and software that can be quite expensive (>>$1,000) [14]. However, the most critical concern pertains to the suitability of the teacher audio quality for ASR. Wang, et al. [30] are agnostic to this issue because they are mainly concerned with a coarse-grained analysis of instructional activities (i.e., lecturing, discussion, and group work), which can be achieved by modeling teacher-student

turn-taking dynamics in a content-free manner [31]. Consequently, they do not report any ASR results. However, in addition to classifying instructional activities, our analysis of dialogic instruction requires identification of questions and classification of questions with respect to authenticity and uptake. The content of teacher speech is needed for this fine-grained level of analysis [12, 27], which is why the Wang approach is unsuitable for our project.

1.3 Proposed Approach

Our core approach emphasizes recording high-quality teacher audio given that the intended application is teacher professional development and that teachers can be individually mic'ed. Further, given that students could not be individually mic'ed, our design focused on simply identifying when students are speaking rather than identifying what they are saying. We expect that certain features of dialogic instruction can be revealed by the sequence of speakers. In particular, we believe that lecturing is more likely to have a pattern of *teacher-teacher-teacher-student* compared to question-answer segments (*teacher-student-teacher-student*). Even within question-answer segments, test recitation might have a pattern similar to *teacher-student-teacher-student*, but an authentic discussion is more likely to have a pattern encompassing transitions between students (i.e., *teacher-student-student-student-teacher)*. We also expect that the duration of student responses is another critical index of question properties in that short student responses should mark test questions whereas elaborated responses are more likely to index authentic teacher questions. Thus, our design is intended to optimize teacher audio as the primary channel. Classroom audio (and subsequently video) is considered to be a secondary channel that is primarily used to contextualize the teacher audio channel.

We identify the following three core analytic tasks motivated by our core approach. By core analytic tasks, we mean that outputs of these fundamental tasks are intended to serve as inputs towards the overall goal of automatically analyzing dialogic instruction as defined above. The three core tasks include:

- *Teacher automatic speech recognition*: Obtain text transcriptions of teacher speech.
- *Teacher utterance segmentation*: Obtain start and end times of teacher spoken utterances.
- *Student utterance segmentation*: Obtain start and end times of student spoken utterances without attempting to identify individual students.

In this paper, we discuss our approach to instrumenting the classroom for data collection suitable for automatic analysis of classroom discourse and report results pertaining to our three core tasks. Although our approach was designed for a specific application (CLASS 5), it should be beneficial for researchers interested in automatic analysis of teacher-focused classroom discourse more broadly.

The remainder of the paper is organized as follows. We discuss the three recording designs that were developed (Section 2) and used to collect data from 38 live classrooms over a two year period (Section 3). The data was analyzed with respect to the three primary tasks of automatic recognition of teacher speech, teacher speech segmentation, and student speech segmentation (Section 4). Plans for incorporating visual information into the design to produce a fourth design are discussed in Section 5.

2. RECORDING DESIGNS

We experimented with a variety of recording designs in order to satisfy the aforementioned requirements and constraints. These include omnidirectional laptop microphones, phase microphone arrays, headset microphones, and pressure-zone-microphones (PZM). We simulated classroom environments in the lab as well as collected classroom data in authentic field environments in order to evaluate these designs. These tests can be organized with respect to three recording designs as discussed below.

2.1 Design 1

2.1.1 Omnidirectional Microphone

We did not believe that basic omnidirectional microphones would be sufficient to record audio of sufficient quality for teacher speech recognition. Nevertheless, we collected a few sessions of classroom audio data with an omnidirectional microphone (built into a laptop) to make sure that we were correct. This was an important option to exclude because omnidirectional microphones on commodity laptops, tablets, and smartphones, if they could work, would allow us to disseminate CLASS5 on these platforms without special hardware. Qualitative evaluations conclusively revealed that this was not the case. Further, ASR using Sphinx 4 [29] and Microsoft Speech SDK 5.1 [19] on teacher speech recorded via an omnidirectional laptop mic produced gibberish. Thus, we determined that special microphone hardware was required to record adequate audio, leading us to Designs 2 and 3.

2.2 Design 2

Design 2 utilized a specialized headset microphone for teacher speech and two first-generation Kinect to record classroom audio consisting of both student and teacher speech. These devices are shown in Figure 1.

Samson 77 Airline Microphone (Teacher audio)

First- generation Kinect (Classroom audio)

Figure 1: Devices in Design 1

2.2.1 Samson Airline Microphone (Teacher Audio)

We selected the Samson 77 Airline microphone system (AH1 Headset Transmitter with CR77 Wireless Receiver), a wireless microphone that is also marketed for aerobics instructors. We performed qualitative evaluations using pre-recorded classroom audio as background noise. These informal analyses revealed that the noise cancellation properties of the Samson were sufficient to overcome classroom background noise even when that noise was played at 2x real levels.

2.2.2 First-generation Kinects (Classroom Audio)

Our design consists of two first-generation Kinect microphone arrays placed in the front corners of the classroom for recording of classroom audio (see Figure 2). Preliminary qualitative testing revealed that Kinect microphones, which are among the least expensive on the market, easily match the quality of the Acoustic Magic Voice Tracker™ array microphone. In addition to recording audio, we can also record positional information using the Kinects. More specifically, each Kinect calculates a "beam" using the difference between the arrival of sound to its microphone elements. The beam indicates the direction of the current sound. By calculating the intersection of the Kinect microphone array beams, i.e., triangulation, we can localize the sound (see Figure 3) within a 3-foot area. We hypothesize that location can serve as a proxy for

identity, because students generally stay in the same seats in a given class.

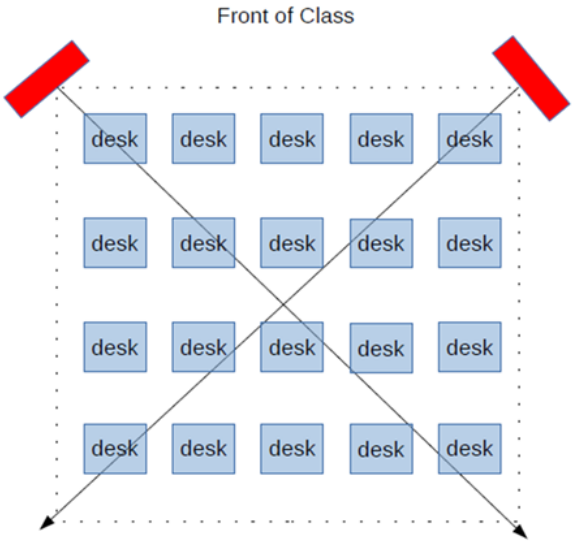

Figure 2. Kinects (shown in red) are aligned to form a coordinate plane, rotated 45 degrees

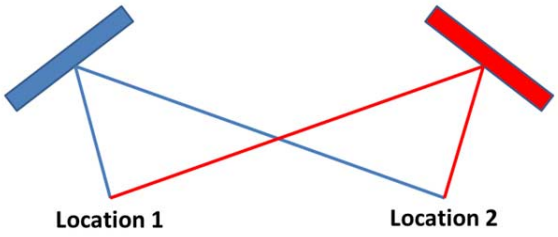

Figure 3. An example of triangulation. At any one time, the red (right) and blue (left) Kinects have a preferred direction, or beam, indicated by the colored lines. When these lines converge, they identify a unique location via triangulation.

2.2.3 Synchronizing Streams

Using the Microsoft Kinect SDK, we created a multithreaded software recording application that synchronized the Kinect streams and teacher audio stream. Each Kinect stream consisted of waveform audio data (i.e., wav files) and Kinect beam data (angle, timestamp, and confidence) whereas the teacher stream consisted of audio data only. All three streams were synchronized on a single laptop with dual USB hubs, which is required when simultaneously recording from two Kinects. Each Kinect performed analog/digital conversion internally. The teacher mic was interfaced using an integrated analog/digital converting XLR to USB cable. The cable utilized stock Windows 7 plug and play drivers. The synchronization software did not have the facility to adjust audio levels, so the automatic gain control (AGC) feature from Windows 7 was used to dynamically adjust volume of the teacher mic.

Synchronization of Kinect beam data was achieved with millisecond precision using hardware-defined timestamps. Specifically, each Kinect produced a sparse list of timestamps, and a dense (aligned) list of pairs was created by filling in each missing value with the preceding value in the respective series. This approach was necessary because the Kinects report beam changes rather than sampling at a fixed interval, and a sound source may change location with respect to the beam of one Kinect without changing its location with respect to the other Kinect. The accuracy and sensitivity of beam pairings was qualitatively verified by recording sound bursts at various locations in a room and matching these locations to the triangulated location derived from the paired beams.

The researchers (called coders), who were trained to perform live coding, started recording by clicking a button on the recording interface. A separate program (CLASS 4.24) was used for live coding of classroom discourse and ran on the same computer as the stream synchronization program. Thus, synchronizing the CLASS codes and data streams (waveforms, position info) simply involved automatic indexing of coded events within the data streams based on recorded start times of the streams (audio data) or time-stamped event data (live coding and position data).

2.2.4 Evaluation of Design 2

Design 2 was tested by collecting several months of audio in authentic classroom environments. We discuss a number of qualitative assessments here; more formal analyses are presented in Section 4.

- The noise cancellation properties of the Samson Airline Microphone were excellent. Therefore, we believe we have found a viable solution to the challenge of recording high-quality teacher speech in authentic classroom conditions.
- The plug-and-play nature of the recording setup, necessitated the use of automatic gain control (AGC). However, this was undesirable as AGC can reduce ASR accuracy by as much as 20% [16].
- Student voice quality recorded from the Kinects was short of what would be required for automatic utterance segmentation.
- The Kinects were wired, so this limited the optimal placement required for accurate sound source localization (see Figure 2). Also, teachers confounded the location information derived from the Kinects by moving about the classroom.
- The software-level synchronization was not perfect because the application was multithreaded. We observed a 0.5 to 2 second discrepancy in the lengths of the resulting audio files.

2.3 Design 3

The evaluation of the data collected with Design 2 provided evidence for the success of using the Samson 77 Airline microphone for recording audio from teachers. We also noted several limitations with the Kinect-based solution to recording classroom audio. Thus, a replacement microphone (a pressure-zone microphone or PZM-30D [6], Figure 4) was added to our classroom recording solution in order to improve classroom recording quality (combined teacher-student speech) without having to resort to mic'ing individual students (and violating a technical constraint).

2.3.1 PZM-30D (Classroom microphone)

A lab based study [5] was conducted to select a suitable classroom microphone for our needs. Several microphones (Perceptron 120, PZM-30D, PZM6D, SM58, Kinect, and Apple Laptop Mic) were tested in an empty classroom with variable conditions. These included either no-noise or high-noise conditions, variable microphone placement (flat on teacher's desk or propped up on chalkboard), and variable live speech (obtained by a speaker walking to four corners of a classroom's student seating area and

speaking from each corner). We also evaluated an omnidirectional non-boundary microphone which acted as a baseline for unacceptable speech quality since we had previously evaluated this style of microphone to record classroom audio (see Section 2.1).

The PZM-30D (see Figure 4) was selected as the best performing mic in these tests. PZMs or pressure zone microphones are omnidirectional boundary microphones that are placed on large surfaces (e.g., a table, wall) and minimize interference from surface reflections. They are designed to record rooms of speech or music while minimizing noise. The tests also indicated that the PZM-30D worked best when placed upright. Typically this included leaning the microphone against a chalkboard, though it can also be placed upright on the teacher's desk or anywhere in the front of the room. Qualitative tests with parallel devices placed on the side walls and floors of an authentic classroom were also conducted, with no notable performance improvements. This is particularly advantageous since most classrooms have a chalkboard (or whiteboard) and a teacher's desk.

Crown PZM-30D (mounted on board)

M-Audio M-Track

Figure 4: Additional devices in Design 3

2.3.2 Stream Synchronization

An audio interface device, an M-Audio M-Track (see Figure 4), was added to the classroom recording solution for stream synchronization. The M-Track has several benefits. First, the teachers' speech and the students' speech are aligned by making each a channel in a stereo recording. In other words the M-Track aligns the channels at the hardware level, rather than the software level, which ensures exact alignment of the two audio streams. This addressed the synchronization problem noted in Design 2. Second, the M-Track simplified the recording setup to a point where an open-source audio recording and editing software (Audacity) was sufficient for recording. Audacity's Timer Record feature, which allows one to specify an exact time to start recording, was used to synchronize the audio with the CLASS 4.24 coding software. Third, in lieu of automatic gain control (which is disabled in Design 3), each microphone has a volume adjustment that researchers/teachers can adjust dynamically to avoid clipping. The adjustment is done based on a live wavform (from Audacity – see Figure 5) being displayed during the first few minutes of the recordings.

2.3.3 Recording Case

Setup complexity was increased with the inclusion of the M-Track and the PZM-30D mic. Individuals needed to be able to switch classrooms between class sessions, which requires a full tear down and setup of audio equipment in a time period as short as five minutes. The solution was to build physical recording case which compactly contained all recording equipment, requiring only that a person place the case near an outlet, plug in a power cable, plug in a USB outlet into a computer, setup the teacher's headset mic, and place the PZM-30D microphone at the front of the classroom.

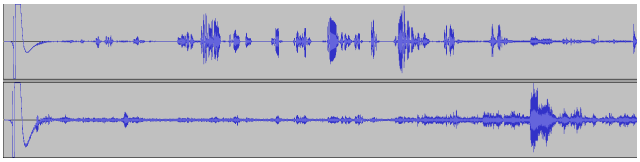

Figure 5. Teacher audio from Samson 77 Airline mic (top) and classroom audio from PZM-30 D mic (bottom)

2.3.4 Evaluation of Design 3

We collected several weeks of classroom audio from authentic classroom environment. Qualitative analyses suggested that the recording quality was acceptable and no major problems were identified. More formal analyses are discussed in Section 4.

3. DATA COLLECTION

We recorded more than 30 hours of data using the recording Designs 2 and 3. Data was collected in rural Wisconsin middle schools during literature, language arts, and civics classes. The audio streams were accompanied by human-coded observations according to the CLASS scheme using the CLASS 4.24 coding software (see Introduction for a description of the coding scheme). Table 1 provides a high-level overview of the two data sets.

Table 1. Overview of data sets collected using proposed recording designs

	Design 2	**Design 3**
District	District A	District B
# Schools	1	5
# Teachers	3	8
# Sessions	21	17
Hours of Data	15.6	16.7
Teacher audio mic	Samson 77 Airline	Samson 77 Airline
Class audio mic	Kinect	PZM-30D

4. DATA ANALYSIS

The overall goal of the project is to develop algorithms to replicate the human-coded CLASS codes with computer-generated codes. In this paper, we report results on three core analytic tasks that need to be addressed in order to achieve the foundational goal. These include: (1) teacher automatic speech recognition; (2) teacher utterance segmentation; and (3) student utterance segmentation. Preliminary results of (1) and (2) using teacher audio collected in Design 2 are reported in [3] and [4], respectively. Here, we increase the scope of the analyses by including teacher audio from Design 3, effectively doubling the sample size and introducing considerable more variability in terms of teachers and class sessions (see Table 1). Preliminary unpublished results on (3) obtained from data collected with Design 3 are presented here.

4.1 Automatic Recognition of Teacher Speech

Automatic speech recognition (ASR) is an important first step in recognizing questions and classifying question types from classroom audio because it enables the application of natural language understanding techniques. However, speech recognition remains a challenging research problem in noisy environments. To

determine the suitability of existing ASR technologies for CLASS 5, we analyzed several out-of-the-box ASR engines that do not require training on speakers and do not require any domain-specific knowledge. We focus on questions asked by teachers because they are the most essential component of dialogic instruction (see Introduction). The goal of the analysis was to evaluate the feasibility of ASR on the current data and to identify which ASR engines are best suited for transcription of teacher speech in classrooms.

4.1.1 Method

We considered four ASR engines: Google Speech [28], Bing Speech [17], AT&T Watson [10], and Kaldi [25]. We also examined Sphinx 4 [29], Microsoft Speech SDK 5.1 [19], and RASR [32] at an early stage of this work, but these were quickly abandoned due to poor performance on our data (see [3] for details). Google Speech, Bing Speech, and AT&T Watson are query-oriented, cloud-based recognition systems primarily intended for web-queries on mobile devices (typically noisy conditions). Kaldi is a speech toolkit written in C++ and is intended for research purposes.

We processed the 1,118 questions recorded across the 38 class sessions by submitting them to each of the ASR engines. We then compared the transcriptions from the engines with human-transcriptions provided by the CLASS coders. Performance metrics were word accuracy (WAcc) and simple word overlap (SWO). WAcc is the complement of the standard ASR metric of word error rate (WER). (WAcc = 1 – WER). WER is calculated by dynamically aligning the ASR engine's hypothesized transcript with the human's transcript and dividing the number of substitutions, insertions, and deletions required to transform the transcript into the hypothesis by the number of words in the transcript. SWO is the number of words that appear in both the hypothesized transcript and the human's transcript divided by the total number of words in the human's transcript. WAcc preserves word order while SWO ignores it. WAcc is bounded on [-∞, 1] while SWO is bounded on [0, 1]. Higher numbers indicate better performance for both metrics.

4.1.2 Results

Table 2 presents descriptives on WAcc and SWO by ASR sorted in descending order of WAcc. This table was produced by first computing the mean WAcc or SWO for each teacher and then computing means and standard deviations across the 11 unique teachers in our sample. A repeated-measures ANOVA with teacher as the unit of analysis and ASR as a four-level factor yielded a significant main effect of ASR on WAcc, $F(3, 30) = 11.2$, $MSE = .013$, $p < .001$. Post-hoc tests indicated the following pattern of significance (at $p < .05$[1]) in the data: Google > Bing > [Kaldi = AT&T]. There was also a significant main effect of ASR on SWO, $F(3, 30) = 24.0$, $MSE = .004$, $p < .001$ with the following pattern in the data: [Google = Kaldi] > Bing > AT&T. Thus, Google Speech outperformed the other engines when word order was considered (WAcc metric), but tied with Kaldi when word order was ignored (SWO metric). Importantly, almost 70% of the words could be correctly recognized using Google Speech — no small feat given the noisy nature of the classroom environment and the conversational speaking style of the teachers.

[1] $p = .057$ the for Bing vs. Kaldi pairwise comparison.

Table 2. Mean speech recognition accuracy across questions (with standard deviations in parentheses)

ASR	Word Accuracy (WAcc)	Simple Word Overlap (SWO)
Google Speech	.655 (.081)	.690 (.075)
Bing Speech	.533 (.133)	.615 (.099)
Kaldi	.439 (.251)	.671 (.116)
AT&T Watson	.391 (.178)	.477 (.126)

We performed a follow-up analysis to study variability in WAcc and SWO as a function of teacher (coded as indicator variables representing each teacher) and speech characteristics (speech rate and verbosity of individual utterances). The analyses preceded by separately regressing WAcc and SWO on the three factors for each ASR engine. Table 3 shows the proportion of variance explained (i.e., Rsq.) in each model. We note that Google and Bing were largely robust to teacher identity and speech characteristics as they explained a mere 3% and 6% of the variability in ASR accuracy, respectively. In contrast, Kaldi and AT&T were severely affected by teacher identity and speech characteristics as these factors explained between 13% to 19% of the variability. Thus, we selected Google and Bing for further analysis as these ASRs were quite accurate and largely invariant to differences in teachers and speech characteristics.

Table 3. Proportion of variance explained (Rsq.)

ASR	Word Accuracy (WAcc)	Simple Word Overlap (SWO)
Google	.031	.028
Bing	.056	.053
Kaldi	.192	.130
AT&T Watson	.141	.126

4.2 Teacher Utterance Segmentation

The goal of this analysis was to convert the teacher audio stream into instances of teacher speech vs. no speech. We developed and validated an utterance detection method and applied it to the teacher audio recorded across the 38 class sessions (see Table 1).

4.2.1 Method

Teacher speech was recorded from a high-quality noise-canceling headset microphone. Thus, we assumed that most sound was voice and that preprocessing using advanced voice activity detection (i.e., detecting presence or absence of speech) or speaker diarization techniques (i.e., segmenting utterances in an audio stream by individual speakers) were not required[2]. Thus, a simple binary procedure was used for utterance detection. Specifically, the amplitude envelope of the teacher's low-pass filtered speech was passed through a threshold function in 20 millisecond increments. Where the amplitude envelope was above threshold, the teacher was considered to be speaking. Where the amplitude envelope was below threshold, the teacher was assumed to not be speaking. Any

[2] We also experimented with an off-the-shelf voice activity detection and speaker diarization algorithm [26] with comparable, if not slightly inferior, results.

time speech was detected, that speech was considered part of an utterance, meaning there was no minimum threshold for how short an utterance could be. Utterances were marked as complete when no speech was detected for 1000 milliseconds (1 second). An example result of this automatic utterance labeling method is depicted in Figure 6.

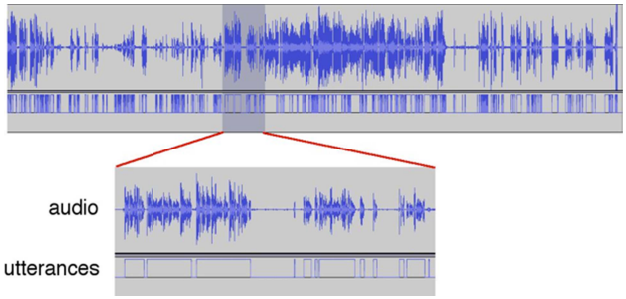

Figure 6. Example of teacher utterance segmentation. An approximately 45-minute class recording (top) with a small portion of the recording enlarged for a detailed view (bottom). The upper track visualizes the .wav form of the audio. The lower track visualizes detected utterances.

The speech delimiter and threshold were both set to be low to ensure all speech was detected, resulting in no known cases of missed speech. This process resulted in 17,298 utterances, which we call *candidate utterances*. An examination of a random subset of these candidate utterances indicated that there were a large number of false alarms. These were mainly attributed to aforementioned low threshold which made it difficult to discriminate background noise from speech. Common examples of background noise picked up by the microphone included voices of students who were being exceptionally loud, sounds from a film or audio clip being played in the classroom, and sounds of teachers' heavy breathing.

A two-step filtering approach was taken to eliminate the false alarms. First, candidate utterances less than 125 milliseconds in length (12% in all) were deemed to be too short to contain meaningful speech and were eliminated. Second, the remaining candidate utterances were submitted through an automatic speech recognizer (Bing Speech) in an effort to identify the false alarms. Bing was selected because it was one of our top performing ASR engines; the other top ASR engine, Google, is also being considered. Bing returns a recognition result and a confidence score for the transcribed utterance. Instances where Bing rejected the utterance or where it returned no transcribed text were considered to be false alarms. After eliminating the false alarms, we were left with a total of 10,426 utterances (60% of the 17,298 candidate utterances).

4.2.2 Validation

A study was conducted to evaluate the aforementioned teacher utterance detection method. A random sample of 1,000 candidate utterances was selected and manually annotated for speech/non-speech. Speech was defined to include all articulations (i.e., "um", "hm", "sh", etc.) in addition to normal spoken segments. Candidate utterances which included noise (i.e., loud students) in addition to teacher speech, were deemed as valid utterances since they contained teacher speech. In total, 59% of the candidate utterances were classified as containing teacher speech, indicating a false alarm rate of 41% prior to discarding utterances based on our filtering method.

Table 4 presents the confusion matrix obtained after the two-step filtering process on the sample of 1,000 candidate utterances. The filtering approach was highly successful, resulting in a kappa of 0.94 (agreement between computer-segmented teacher utterances and human-coded teacher utterances). Precision (96.3%) and recall (98.6%) were excellent, resulting in an F-measure of 97.4%. Thus, the present method was deemed to be sufficiently accurate for automatic teacher utterance segmentation.

Table 4. Confusion matrix for teacher utterance segmentation

	Predicted	
Actual	*Speech*	*Non-Speech*
Speech	0.963 (hit)	0.037 (false alarm)
Non-Speech	0.020 (miss)	0.980 (correct rejection)

4.3 Student Utterance Segmentation

The primary goal of student utterance segmentation is to identify student utterances from the classroom audio stream. Although high-quality speech recognition of student utterances would be ideal, our working hypothesis is that identification of student utterances and their surface properties, like duration, may be useful for dialogic question classification even when audio quality, and thus speech recognition, is poor (see Introduction). Moreover since the query-oriented ASR systems described in Section 4.1 require short duration inputs, student utterance segmentation is a necessary preprocessing step.

4.3.1 Method

The key challenge in student utterance segmentation is that the PZM channel records all classroom audio. Because the PZM mic is placed in the front of the class, the teacher's speech signal is as strong, or stronger, than the student signal. In contrast, the teacher headset microphone, due to its directional properties, proximity to the teacher's mouth, and filtering capabilities, produces a teacher speech signal that is substantially stronger than students' speech (see Figure 5). For these reasons, the relatively simple amplitude envelope followed by our two-step filtering approach used for teacher utterance segmentation (see Section 4.2.1) is unlikely to be applicable for student speech segmentation; qualitative tests quickly confirmed this suspicion.

To address this issue, we developed an alternative approach for student speech segmentation. This approach has two steps. First, we used an off-the-shelf diarizer - the LIUM diarizer [26] - to generate *candidate student segments*. The LIUM diarizer is an open source diarization system created for broadcast news. It performs diarization by extracting acoustic features, segmenting, and clustering, typically in a pipeline with interleaving and clustering stages. LIUM can perform gender identification, speaker identification, and speech/non-speech identification. However, the classroom audio under discussion is sufficiently noisy that only speech/non-speech discrimination appears to be feasible.

Our initial investigations revealed that LIUM critically depends on a few manually set parameters for minimum speech and silence lengths along with transition probability parameters for speech/music/silence segmentation. The parameters influence both the size of resulting segments and the filtration of noisy segments, such that poor parameter choices can either group multiple utterances together and filter away smaller segments or fragment single utterances into many small segments and fail to filter small segments (a classic precision/recall tradeoff). Fortunately, preliminary analyses revealed that the internal parameters, once set

for a particular recording context, appeared to be fairly stable for other recordings in that context. However, this does need to be verified with more systematic analyses.

As mentioned, the resulting candidate segments from LIUM contain both teacher and student utterances, since the PZM records the entire classroom. Thus, the second step involved using the high-fidelity teacher segments to filter teacher segments from the candidate segments (student + teacher) returned by LIUM. The primary assumption of our filtering approach is that overlapping speech between teacher and student is rare, and when it occurs, we can assume that the teacher's speech has priority. Using this assumption we can filter all LIUM-obtained candidate segments that intersect the high-fidelity teacher segments by simply treating those as teacher speech. However, this naïve approach turned out to be too strict because it eliminated student segments on the boundary with teacher segments, as might occur when student segments have been marginally lengthened due to LIUM error.

Therefore, we adopted a slightly more complex approach shown in Figure 7. We used the high-fidelity teacher segments (from dedicated teacher channel) to re-segment LIUM-candidate segments. If a candidate segment intersected a teacher segment, the overlapping portion was removed, either deleting, shortening, or splitting the LIUM segment depending on the nature of the overlap. Since multiple teacher segments potentially overlap with a given candidate segment, and since the re-segmentation process itself potentially generates new segments, the re-segmentation process was recursive and terminated when every teacher segment had been applied to every candidate segment. We then merged all segments from the same speaker (remaining segments after the filtering process are now assumed to be student segments) that have less than 1 second of silence between them. After this step, segments shorter than 250 milliseconds are removed. These two parameters (1 second and 250 ms) have been fixed in these analyses but are tunable.

4.3.2 Validation

We conducted a study to evaluate the student utterance segmentation procedure. We randomly selected a subset of teacher questions and then manually coded all audio between the start of each question and the start of the next question as teacher speech, student speech, or silence, resulting in 298 annotated segments spanning 15 class sessions. We focused our analyses on student speech or silence immediately after teacher questions because these are the most interesting moments from the perspective of dialogic instruction. Thus, the 298 manually annotated segmented constitute our "gold standard."

Next, the entire PZM audio files from which these segments were drawn were submitted for teacher segmentation (using the method specified in Section 4.2.1) and to LIUM, and then the resulting segmentations from these methods were merged as specified in Section 4.3.1. For each annotated segment, the corresponding time interval in the merged segmentations was located and the interval classification was compared to the manual segment annotations. In order to take the duration of the segments into account, comparisons were done at the $1/100^{th}$ of a second level. This allowed us to consider segmentation accuracy as a standard classification problem and use precision/recall metrics, i.e. each $1/100^{th}$ of a second represented one opportunity for the "classifier" to identify whether the classroom audio reflected teacher speech, student speech, or silence (N = 99,290 in all).

The accuracy of the student utterance segmentation at the $1/100^{th}$ of a second level was fairly high, with precision of 79.6%, recall of 55.8%, and F-measure of 65.6%. The misclassifications of student segments to teacher or silence were approximately evenly distributed, suggesting that using the teacher segmentation to guide the student segmentation did not bias student segmentation according to the false alarms or misses of teacher segmentation. It appears that performance was largely limited by student utterance detection accuracy (F-measure of 65.6%). Results can be considered to be moderate given the noisy environment and technical constraints of not being able to mic individual students.

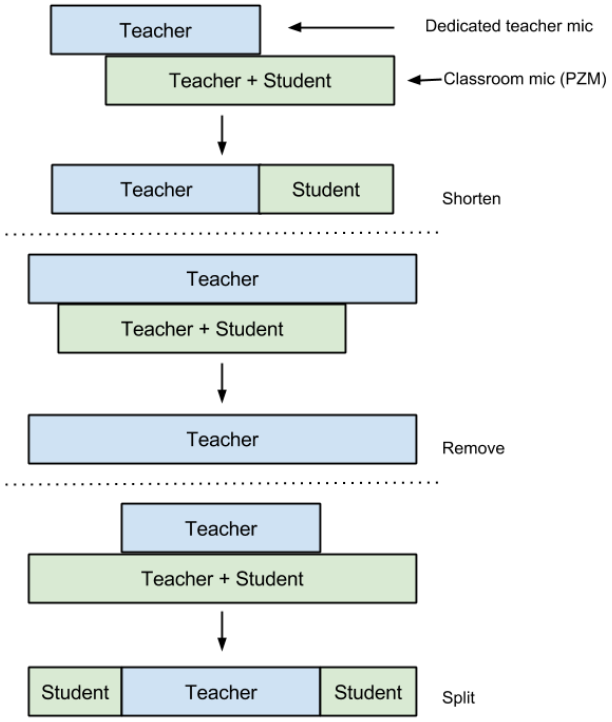

Figure 7. Student utterance segmentation. Overlap with teacher segments identified from the Airline mic shorten (top), remove (middle), or split (bottom) candidate segments recorded with the PZM mic to create student segments.

5. DESIGN 4 (Visual Information)

The key limitation in our designs to date boils down to physics (the least noisy signal must be collected close to the source) and to the practicalities of modern classrooms (instrumenting each student with a microphone is impractical). One possibility is to identify a set of verbose students and individually mic these students or to mic small groups of students. These options offer a middle ground to the current dichotomy of mic'ing individual students vs. none of the students and need to be tested more extensively. It is also worth considering what other kinds of sensors might be applicable to this problem.

In an early design we experimented with Kinects but did not use their video capability because of privacy concerns. It may be possible, however, to use the Kinects to collect visual data in a manner that would not raise privacy concerns. For example, the first-generation Kinects have the capability of collecting depth information. This depth information is sufficiently coarse that facial identification is obfuscated if not impossible. The skeleton tracking capabilities of the first-generation Kinect, which are only functional for single individuals, have been greatly expanded and made functional for small groups of individuals in the second generation Kinect. The second-generation Kinect can detect whether the eyes or mouth are open, if a person is looking away, and if the mouth has moved, for up to six people at a time [18].

In addition to using the mouth open/closed/moved information to better determine whether a student is speaking (by integrating this information with audio information collected by the microphone and PZM), the measures mentioned above could be used to estimate the students' levels of engagement. After all, the dialogic properties we are investigating are to some extent a linguistic measure of joint attention and engagement. It is an open question as to whether other measures, such as the degree to which students in the class are all looking at the teacher, are correlated with or even more predictive than linguistic measures on engagement.

In light of the aforementioned discussion, we are considering adding second-generation Kinects to Design 3, effectively producing Design 4. Design 4 will consist of the following devices and will yield the following data streams (in parentheses).

- CLASS 4.24 coding program (live codes of classroom discourse provided by humans)
- 1 Samson 77 Airline Microphone (teacher audio)
- 1 Crown PZM-30D (classroom audio)
- 2 second-generation Microsoft Kinects placed similar to Figure 2 (classroom audio, speaker position information, body tracking, and face tracking)

Synchronization of these multiple streams will likely pose a challenge. Synchronization of items 1-3 have already been completed as part of Design 3, but synchronization with the Kinects will need to be worked out, especially given previous difficulties involving synchronization of Kinects (see Section 2.2.4).

It should also be noted that the second-generation Kinects can only track up to six students simultaneously, which makes tracking the entire class impractical in terms of cabling around the classroom. This suggests that only the first and perhaps part of the second row can be successfully tracked with the proposed two Kinects. However, this limitation may represent the most pragmatic balance between the capabilities of modern sensors and the practicalities of classrooms. Moreover, there is some evidence to suggest that most question events involve the front of the class: in a within-subject study, students randomly assigned to the front asked more questions than when they were assigned to the back of the class [15]. Therefore the inability of the Kinects to track more than the front of the class may not be such a limitation after all.

6. DISCUSSION

The present project was concerned with the development of CLASS 5 – a program to autonomously code classroom discourse with respect to instructional activities (e.g., lecturing, discussion) and dialogic instruction quality (e.g., authenticity of questions; uptake). We developed and tested three recording designs in 38 live classrooms in multiple schools in the U.S. Our approach was motivated by a set of 12 technical requirements and constraints (see Section 1.1). The current most successful design (Design 3) met or exceeded the aforementioned requirements/constraints with respect to: usability, economic cost (around $875), human cost, scalability, non-intrusiveness, audio information, teacher audio quality, stream synchronization, and identity resolution. It was moderately successful in terms of flexibility and student audio quality, though the former needs to be studied more conclusively.

Analyses indicated that the data collected via Design 3 was suitable for our three core data analytic tasks: automatic speech recognition on teacher speech (word recognition rate of 66% and word overlap rate of 69%); teacher utterance segmentation (F-measure of 97%); and student utterance segmentation (F-measure of 66%). We consider these results to be promising due to the challenging nature of the classroom context in that we are instrumenting the classrooms with head-worn teacher microphones and boundary microphones, giving us a mixture of relatively clean teacher audio and extremely noisy student audio.

Importantly, the outputs of these basic tasks have been successfully applied towards the broader goal of automatic analysis of classroom discourse. As a concrete example, Blanchard, et al. [4] studied the possibility of automatically detecting question-answer segments (an example of instructional activity identification) in live classrooms based solely on audio recordings of teacher speech. Our approach had two steps. First, teacher utterances were automatically detected from the audio stream using the method specified in Section 4.2.1. Second, supervised classifiers were trained on speech-silence patterns derived from the segmented teacher utterances. We were able to detect Q&A segments in windows of 90 seconds with an AUC (Area Under the Receiver Operating Characteristic Curve) of 0.78 and in a manner that generalizes to new classes (i.e., using class-level cross validation). Similar results for identifying other instructional activities (e.g., small group work, lecturing), question detection, and question property classification, are underway and preliminary results are promising. We also have been working on using the data collected with these designs to identify questions in utterances and further identify questions with authenticity and uptake.

Despite its success, Design 3 was limited by its inability to record visual information (mainly due to privacy requirements). To address this, we proposed Design 4 which extends Design 3 to incorporate visual information from two second-generation Microsoft Kinects. We expect that the additional multimodal information available in Design 4 will help us to further improve the core task of student utterance segmentation. This, in turn, will be essential for more accurately coding classroom discourse, thereby providing valuable information to researchers, teachers, teacher educators, and professional development personnel. This information can be used to help teachers improve on their use of dialogic instruction techniques with the overall goal of improving student engagement and achievement.

7. ACKNOWLEDGMENTS

We acknowledge Mike Brady and Martin Nystrand for their contributions. 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.

8. 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), 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), 685-730.

[3] 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. In Conati, C.,

Heffernan, N., Mitrovic, A. and Verdejo, M.F. eds. *Artificial Intelligence in Education*, Springer-Verlag, Berlin Heidelberg.

[4] 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. In Romero, C., Pechenizkiy, M., Boticario, J. and Santos, O. eds. *Proceedings of the 8th International Conference on Educational Data Mining (EDM 2015)*, International Educational Data Mining Society.

[5] Brady, M.C., D'Mello, S., Blanchard, N., Olney, A. and Nystrand, M. Year. Evaluating microphones and microphone placement for signal processing and automatic speech recognition of teacher-student dialog. In *168th meeting of the Acoustical Society of America*, (Indianapolis, Indiana, 2014), 2215-2215.

[6] Crown. 2006. PZM-30D PZM-60D.

[7] 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., Boulder, CO.

[8] Gamoran, A. and Nystrand, M. 1992. Taking students seriously. In Newmann, F.M. ed. *Student engagement and achievement in american secondary schools*, Teachers College Press, New York, NY.

[9] Gates. 2013. Ensuring Fair and Reliable Measures of Effective Teaching: Culminating Findings from the MET Project's Three-Year Study, Bill & Melinda Gates Foundation.

[10] Goffin, V., Allauzen, C., Bocchieri, E., Hakkani-Tür, D., Ljolje, A., Parthasarathy, S., Rahim, M.G., Riccardi, G. and Saraclar, M. Year. The AT&T WATSON Speech Recognizer. In *International Conference on Acoustics, Speech and Signal Processing*, (2005), 1033-1036.

[11] Goldman, R., Pea, R., Barron, B. and Derry, S.J. (eds.). *Video research in the learning sciences*. Erlbaum, Mahwah, NJ.

[12] Graesser, A. and Person, N. 1994. Question asking during tutoring. *American Education Research Journal*, *31* (1), 104-137.

[13] Kelly, S. 2008. Race, social class, and student engagement in middle school English classrooms. *Social Science Research*, *37* (2), 434-448.

[14] LENA. 2015. LENA Research Foundation.

[15] Marx, A., Fuhrer, U. and Hartig, T. 1999. Effects of classroom seating arrangements on children's question-asking. *Learning Environments Research*, *2* (3), 249-263.

[16] Microsoft. 2010. Best Practices for Enabling Voice Recognition, Microsoft.

[17] Microsoft. 2014. The Bing Speech Recognition Control

[18] Microsoft. 2015. Kinect for Windows SDK MSDN.

[19] Microsoft. 2014. Speech SDK 5.1.

[20] NCES. 2015. Digest of Education Statistics, 2013, U.S. Department of Education, National Center for Education Statistics, Washington, DC.

[21] Nystrand, M. 2004. Classroom Language Assessment System (CLASS) 4.24, University of Wisconsin–Madison, Madison, WI.

[22] Nystrand, M. 1997. *Opening Dialogue: Understanding the Dynamics of Language and Learning in the English Classroom. Language and Literacy Series*. Teachers College Press, New York, NY.

[23] Nystrand, M. and Gamoran, A. 1991. Instructional discourse, student engagement, and literature achievement. *Research in the Teaching of English*, *25* (3), 261-290.

[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), 135-198.

[25] Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlíček, P., Qian, Y. and Schwarz, P. Year. The Kaldi speech recognition toolkit. In *IEEE Workshop on Automatic Speech Recognition and Understanding*, (2011).

[26] Rouvier, M., Dupuy, G., Gay, P., Khoury, E., Merlin, T. and Meignier, S. Year. An open-source state-of-the-art toolbox for broadcast news diarization. In *Interspeech*, (2013).

[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. In Stamper, J., Pardos, Z., Mavrikis, M. and McLaren, B.M. eds. *Proceedings of the 7th International Conference on Educational Data Mining (EDM 2014)* International Educational Data Mining Society.

[28] Schalkwyk, J., Beeferman, D., Beaufays, F., Byrne, B., Chelba, C., Cohen, M., Kamvar, M. and Strope, B. 2010. "Your Word is my Command": Google Search by Voice: A Case Study. In Neustein, A. ed. *Advances in Speech Recognition: Mobile Environments, Call Centers, and Clinics*, Springer US, New York, NY.

[29] Walker, W., Lamere, P., Kwok, P., Raj, B., Singh, R., Gouvea, E., Wolf, P. and Woelfel, J. 2004. Sphinx-4: A flexible open source framework for speech recognition, Sun Microsystems, Inc, Mountain View, CA, USA.

[30] Wang, Z., Miller, K. and Cortina, K. 2013. Using the LENA in Teacher Training: Promoting Student Involement through automated feedback. *Unterrichtswissenschaft*, *4*, 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* (1), 115-123.

[32] Wiesler, S., Richard, A., Golik, P., Schluter, R. and Ney, H. 2014. RASR/NN: The RWTH neural network toolkit for speech recognition. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* IEEE, Washington, DC.