Challenges and Feasibility of Automatic Speech Recognition for Modeling Student Collaborative Discourse in Classrooms

Rosy Southwell, Samuel Pugh, E. Margaret Perkoff, Charis Clevenger, Jeffrey B. Bush,
Rachel Lieber, Wayne Ward, Peter Foltz, Sidney D'Mello
Institute of Cognitive Science
University of Colorado Boulder
Boulder, CO 80309
roso8920@colorado.edu

ABSTRACT

Automatic speech recognition (ASR) has considerable potential to model aspects of classroom discourse with the goals of automated assessment, feedback, and instructional support. However, modeling student talk is besieged by numerous challenges including a lack of data for child speech, low signal to noise ratio, speech disfluencies, and multiparty chatter. This raises the question as to whether contemporary ASR systems, which are benchmarked on adult speech in idealized conditions, can be used to transcribe child speech in classroom settings. To address this question, we collected a dataset of 32 audio recordings of 30 middle-school students engaged in small group work (dyads, triads and tetrads) in authentic classroom settings. The audio was sampled, segmented, and transcribed by humans as well as three ASR engines (Google, Rev.ai, IBM Watson). Whereas all three ASRs had high word error rates, these mainly consisted of deletion errors. Further, Google successfully transcribed a greater proportion of utterances than the other two, but with more word substitutions; insertions were low across the board. ASR accuracy was robust to different speakers and recording idiosyncrasies evidenced by <5% of variance in error rates attributable to the student and recording session. We found that ASR errors had a larger negative effect on downstream natural language processing tasks at the word, phrase, and semantic levels rather than at the discourse level. Our findings indicate that ASR can be used to extract meaningful information from noisy classroom speech and might be more suitable for applications that require higher precision but are tolerant of lower recall.

Keywords

automatic speech recognition, collaborative problem solving, classroom speech, natural language understanding

1. INTRODUCTION

Students learn by telling and doing. Indeed, decades of educational research has converged on one (among several) perspectives of learning as a social and collaborative activity [8, 89, 97]. Effective collaborative learning (CL) activities give students the opportunity to work together towards a common goal, share their ideas and build upon the ideas of others, negotiate strategies, monitor execution of plans, and reflect on outcomes [17, 28, 33, 39, 72, 75]. Thus, the benefits of CL are manifested not only in the acquisition of domain knowledge [86], but also in the development of essential 21st century skills such as collaborative problem solving and critical thinking [27, 29].

Despite a strong consensus on the value and merits of CL, its widespread implementation in contemporary classrooms is limited. A key factor limiting its adoption is that it is extremely challenging for teachers to effectively orchestrate rich CL activities in their classes. To support successful CL, teachers must monitor group progress on time-sensitive activities, provide guidance and help when students get stuck and risk disengagement, and ensure that students engage in productive knowledge-building conversations, all while ensuring that classroom norms for respectful discourse are maintained [71, 88]. To complicate things further, teachers must perform these demanding activities simultaneously across multiple (often 5-10) groups – a daunting assignment. Can intelligent systems, which unlike teachers, are able to be omnipresent across multiple student groups, enhance teachers' ability to scaffold rich CL experiences for all their students?

One exciting possibility is to design systems capable of natural language understanding (NLU) to support CL in student groups. Indeed, the linguistic content of discourse during CL is considered the "gold mine of information" on how students acquire knowledge and skills [32, 73]. However, despite an extensive body of research demonstrating the utility of other modalities (e.g., body movement, gesture, eye-gaze, paralinguistics, see review [62] for automatically analyzing collaboration, an automated approach for capturing, transcribing, and analyzing student speech during face-to-face CL in the classroom has yet to be developed. Most language-based approaches to date thereby rely on typed transcripts from chats (or human-transcribed speech) to analyze and support collaborative discourse [21, 30, 52, 76].

At the heart of this challenge lies an extremely difficult technical hurdle: using automatic speech recognition (ASR) to obtain accurate (or even serviceable) transcriptions of student discourse in noisy, real-world classrooms. This endeavor is complicated by multiple compounding challenges. Namely, with upwards of 20-30 students in a typical US classroom [57] with multiple student groups simultaneously engaged in CL activities, speech signals are obfuscated by background chatter and ambient noise. In addition, ASR systems already have difficulty recognizing children's speech (even in ideal, noise-free environments), as they tend to speak less clearly than adults [46]. In fact, even the basic acoustic characteristics of children's voices and language use [25, 46], differ from adults (on whose voices most ASR systems are trained), resulting in a degradation in performance when these systems are applied to children's speech [70]. Multiparty speech recognition is another challenge for ASR [12, 65], where utterances - from an unknown number of unique speakers - may overlap, whereas ASR systems are generally trained on audio where speakers have already been separated.

Despite these challenges, pursuing technologies capable of automatically capturing and analyzing student speech during face-to-face CL in authentic school environments is an important avenue of research. These technologies have the potential to significantly improve orchestration and support of CL [73], whether by providing teachers with feedback on progress of student groups (e.g., via a teacher dashboard [87]), or enabling real-time interventions to guide groups of learners towards equitable and productive collaboration.

In this paper we take a first step towards understanding the feasibility and challenges of automatically analyzing student speech in classrooms. Specifically, we investigate: (1) patterns of errors in widely used commercial ASR systems for transcribing student discourse in authentic collaborative learning settings; and (2) the influence of ASR errors on downstream natural language understanding tasks at the word, phrase, semantic, and discourse levels. In doing so, we take an important step towards deploying speech-based collaborative learning technologies in classrooms.

1.1 Background and Related works

There is a large body of research on analyzing student- and teacher-classroom discourse [10, 54], so to keep scope manageable we focus on the automatic analysis of student speech and classroom speech.

1.1.1 Challenges with child speech recognition

Speech recognition in children is a well-documented challenge, with recognition accuracy substantially lower for children's speech than adult's [25, 61, 74]. Yet, both commercially-available and research ASR systems are generally trained with clean audio data from adult speakers, with one speaker per utterance, and often reading from a script, which perform substantially worse on realistic, spontaneous speech [83]. These systems do not easily generalize to child speech where vocal characteristics such as higher fundamental and formant frequency and greater variability in pitch, and linguistic factors such as disfluency rate, differ between children and adults, as well as changing as children mature [25, 46]. In an analysis of Google, Bing and Nuance ASR systems, [70] found that age significantly impacted performance for all ASRs except Google. Further, accented speech of non-native speakers impacts ASR performance, as articulation and pronunciation differ from the training data [19]. The classroom setting provides an additional challenge. Howard et al. [34] reported that the typical classroom signal-to-noise ratios range from −7 dB to +5 dB, further impeding ASR [95]. Finally, microphone placement impacts recognition - the further the speaker from the microphone, the greater the impact of reverberation and other signal degradation on ASR [23, 56].

1.1.2 Child speech recognition in controlled learning environments

Numerous educational applications which use ASR on children's speech have been developed, albeit outside of the hustle and bustle of the classroom. One strand of research uses ASR as part of automatic reading tutors for young children learning to read aloud from text. Here, the reference (ground-truth) transcript is available in the form of reading materials, and by comparing this to the ASR output, pronunciation errors can be identified and fed back to the student or their teacher [3, 51, 60, 66]. Generally, these systems are used in a quiet environment such as a library [66], and in all cases are designed with the expectation that only a single speaker is reading at a time. Another application of ASR is in conversational tutors, where both speech recognition and language generation are combined in a real-time system. One example is My Science Tutor (MyST [92]) which supports one-on-one and small-group science learning [13]. The MyST ASR system was trained using a dataset of elementary school students, and achieves a word error rate (WER) of 0.30 (about 70% accuracy) on a reduced vocabulary of ~6000 words. However tutoring sessions did not take place in the main classroom, and users wore a headset, both of which avoided some of the key challenges of classroom ASR. Online learning environments also simplify the collection of clean, speaker-separated speech recordings, and several examples exist of automated analysis of student-teacher dialog starting from ASR transcripts [47, 94].

1.1.3 Automated analysis of teacher speech in the classroom

Recent advances in ASR make the prospect of sufficiently accurate transcription of speech in the classroom a possibility. Most of the ASR literature focuses on adult speech, and this is mirrored in the availability of commercially available, cloud-based ASR APIs, (for examples, see [20] but see [15] for a child-tailored ASR service). As a result, most automated approaches have focused on analyzing teacher speech with varying degrees of automation including ASR-only [9, 37, 38, 42, 81, 96], human transcripts [82], or a combination of both [7]. There are also differences in the depth of the construct being modeled. For example, Zylich and Whitehill [96] recently aimed to automatically detect 21 key phrases (e.g., "good job") in teacher talk from audio, but stopped short of measuring pertinent discourse constructs. In contrast, Kelly et al. [42] and Jensen et al. [37, 38] developed fully automated approaches to model five features of discourse: questions (vs. statements), authentic (open-ended), instructional utterances, elaborated evaluations, cognitive level, goal specificity, and presence of disciplinary terms.

One advantage of focusing on teachers is that it is easier to affix high-quality microphones on a single teacher than an entire classroom of students. For example, the Kelly and Jensen studies used a unidirectional, noise-canceling microphone with cardioid pickup pattern which is most sensitive to sounds from the front of the mic, thereby canceling background noise [37, 38, 42]. Despite a high-quality mic, classroom ASR is still challenging due to background noise, multidisciplinary chatter, dialectical variations, and so on. To this point, [5] and [18] compared several ASR engines for accuracy in transcribing teacher speech recorded in authentic classrooms. These two studies tested 7 ASRs yielding word error rates ranging from .31 to 1.00.

It is important that these studies are replicated due to the rapid advancement in ASR technologies each year. For example, using the same microphone and ASR engine on similar classroom data, Jensen et al. [37] obtained a major reduction in error (from 44% WER to 28%) in 2020 compared to Blanchard's (2015) study [5].

1.1.4 Automated analysis of student speech in classrooms

Examples of automated analysis of classroom audio focused on student speech are rare, as justified by the many acoustic and linguistic challenges inherent in the full pipeline from recording speech, to transcription in the context of overlapping, non-adult speakers in a noisy environment, to extracting meaning from language patterns of students still undergoing linguistic development. Nevertheless, several recent works have utilized non-specialized commercial ASR services for child speech with promising results - demonstrating that ASR transcriptions can be used to derive useful downstream measures despite very high WER [64, 84].

To our knowledge, the only example where ASR is used to transcribe conversations among students as input to an NLP model is in the context of a collaborative problem solving (CPS) study conducted in both the classroom and lab [64]. Here, students aged 12-15 participated in two CPS activities in math and physics. Participants wore headsets with microphones and completed the task (in dyads) over Zoom from a shared computer lab at the school, or for a subset of participants, in a laboratory. Captured speech was manually segmented into utterances, then transcribed using the IBM Watson speech-to-text service [36]. Performance degradation attributable to the classroom environment was evident, with a word error rate (WER) of 0.78 in the classroom, meaning only 22% of human-transcribed words were correctly transcribed, as compared to a WER of 0.54 for dyads recorded in the laboratory.

ASR has also been used to capture classroom conversation in preschool children. Lileikyte et al. [48] used LENA's wearable audio recorders, which are designed for capturing speech in young children, to train an ASR with custom acoustic and language models using data augmentation, obtaining a WER of 0.64 on spontaneous conversation in 2–5-year-old children. Using the same wearable devices in preschoolers, Tao et al. [84] ran audio through Google Cloud ASR [26] and used the transcripts to derive network representations of groups in social interactions based on word count vector similarity between utterances, though ASR accuracy is not reported. Further, the use of LENA is cost-prohibitive, with pricing in the thousands of dollars, which is infeasible at scale.

Beyond these examples, speech analysis in the classroom is limited to extraction of non-linguistic (i.e., acoustic/prosodic) features, which nevertheless show promise for classification of discourse categories [6, 40, 91], speaker identification [84] and diarization to identify speaker turns [49, 53].

1.1.5 Is perfect ASR needed?

As reviewed above, ASR in the classroom is beset by many challenges, especially for analyzing student speech. However, the goal of many applications is not to obtain perfect transcripts of speech, but to use the transcripts for downstream NLU tasks relevant to education (e.g., assessment, feedback, intervention). Indeed previous research has indicated that useful information can be obtained from imperfect transcripts. Pugh et al. [64] found that using ASR instead of human transcripts led to only a 14% decrease in classifier performance (still significantly above chance) despite a WER of 0.78. Outside the classroom, Stewart et al. [78] reported a mere 4.2% decrease in accuracy for classifying collaborative skills using ASR versus human transcripts. Indeed, the question of robustness of models of team performance to simulated ASR errors was addressed by [22], with even a WER of 57% only decreasing classifier performance by 20% relative to perfect transcription. The authors suggest that the constrained, contextualized nature of conversation makes discourse-level NLP models robust to modifications of individual words.

Of course, there is likely an upper limit to errors beyond which the signal to noise ratio is too low to be useful, a likely possibility for analyzing multiparty collaborative child speech in the classroom. This raises the questions of whether it is feasible to obtain meaningful information on student collaborative discourse despite noisy ASR and to what extent do ASR errors impact the meaning conveyed in an utterance and how does this impact downstream NLU tasks.

1.2 Current Study, Contribution, & Novelty

In this study, we take an important first step towards the automated analysis of student collaborative discourse in noisy, authentic classrooms. We compare a variety of commercially available ASR systems on both speech to text transcription, and we investigate the influence of ASR errors on downstream NLU tasks using a novel dataset of audio recordings from real-world middle-school classrooms where multiple student groups are engaged in CL.

Specifically, we quantify ASR performance in terms of traditional evaluation metrics (e.g., Word Error Rate [WER]), and investigate the types of speech recognition errors encountered (e.g., substitutions, deletions). Further, we seek an understanding of the sources of variability in ASR errors at the level of the utterance, student, and session by systematically sampling students across multiple recording contexts (i.e., across different lessons, student groups, and days). This information can provide insights into potential disparities of ASR systems, which may have unequal impacts on individual student outcomes when used as inputs to downstream applications. To this point, we also compare ASR errors and their influence on downstream NLU applications (e.g., semantic similarity of transcripts [43], recognition of task-relevant content words, assessing collaboration skills) to probe the feasibility of using automated transcripts for NLU-based CL analytics in the classroom.

To our knowledge, this is the first attempt to systematically analyze automated transcriptions of face-to-face student collaborative discourse in a real K-12 school environment. Although other studies use ASR as input to language-based models of classroom discourse, the majority of these focus on teacher speech [5, 9, 14, 37, 38] or collaborative problem solving in adult undergraduates [63, 78, 79]. We also use inexpensive, commercially available microphones placed on the tabletop, each capturing speech from 2-4 students, which allows us to expose the challenges of capturing real-world classroom audio where multiple speakers are intermixed in a single-channel recording with additional impacts of reverberation and background noise. This contrasts with prior studies analyzing classroom audio, which mostly employ individual microphones to isolate speech [5, 37, 47, 48, 64]. Also, we use data collected in the context of a live, face-to-face discourse rather than an online learning environment [47, 94]. The choice to use table-top mics rather than individual noise-canceling lapel microphones or headsets is motivated both by practicality and cost considerations, and by the concern that individually miking students would feel intrusive and even impede collaboration.

Finally, with respect to scope, we focus on widely available commercial ASR services in lieu of customized ASR systems with acoustic and language models trained on our target demographic and data. This may disadvantage speech recognition performance, however using publicly available ASR providers is desirable for practical reasons including the simplicity of integration due to a well-documented API, and the likely continuation of updates to the model in the future. We also don't seek to improve or engineer better performance out of these systems in the current work because the goal is to establish baseline performance of out-the-box ASR systems on the difficult task of analyzing child collaborative talk in noisy classrooms.

2. METHODS

2.1 Data Collection

The data was collected as part of a larger project involving a Research-Practice Partnership [41] focused on using co-design and professional learning to support the use of programmable sensor technology and computational thinking for authentic inquiry in middle school science and STEM classrooms [4]. We analyzed audio and video data from one participating U.S. public middle school teacher in this work.

2.1.1 Learning Context: Sensor Immersion

Participating teachers implemented a multi-day curriculum unit called Sensor Immersion that focuses on students working collaboratively to understand how to program and wire sensors to collect data about their local environments, empowering students to be data producers [31] and answer questions that they find personally meaningful and relevant. The Sensor Immersion curriculum uses an interactive data display called the Data Sensor Hub (DaSH [11]; Figure 1) as an anchoring phenomenon [24]. Students explore the system, create scientific models and learn to replicate its functionality in the context of their own investigations. Along the way, students develop a program that can control a variety of physical sensors including a sound sensor, moisture sensor, and an environmental sensor.

Sensor Immersion is broken down into five lessons, each of which can span multiple days. Lesson 1 focuses on question generation and modeling. Throughout the following lessons, students work to answer their questions about how the DaSH works. To do so they learn to program and wire the sensors, working in pairs doing a pair-programming task using MakeCode block programming (Figure 2). Students gradually build on their understanding of programming and sensors by working together to program and wire one sensor and eventually building and programming a sensor system to answer questions about a personally meaningful phenomenon. Opportunities for small-group collaboration around these sensors and their programming are designed into each lesson.

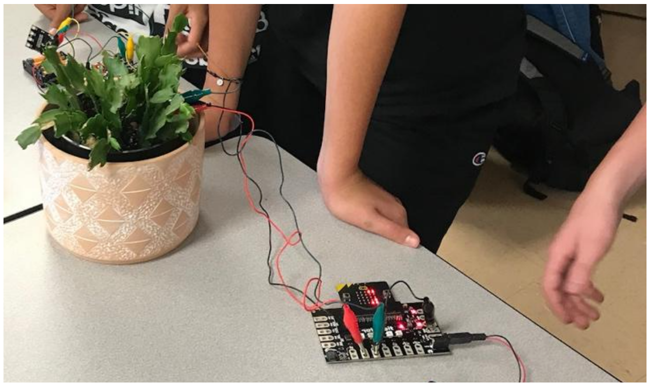
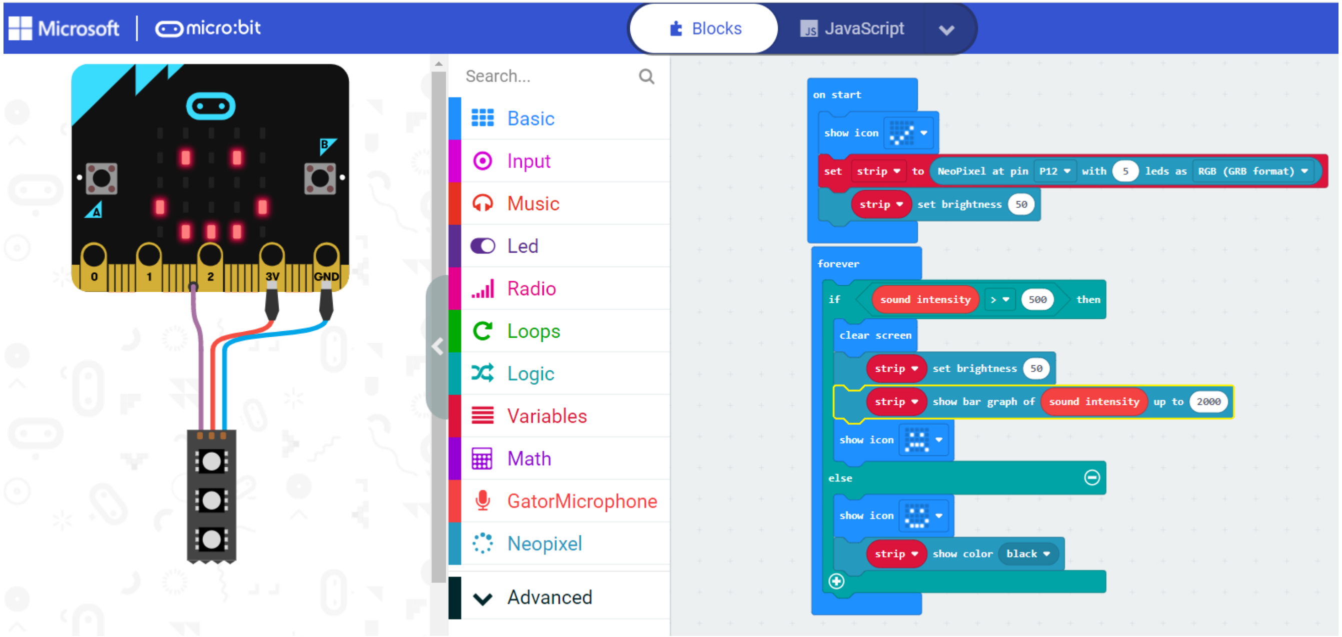


Figure 1. Close-up of the DaSH system which links sensors to the computing interface. Various sensors can be wired to the system to measure local environmental conditions such as soil moisture levels (pictured), CO2, humidity, temperature and ambient room noise.

2.1.2 Participants

The data sample included 30 students from 4 cohorts taught by a single teacher in a suburban school district in the US. All procedures were approved by designated Institutional Reseearch Boards and data were only collected from students who provided both personal assent and their parent's signed consent forms. Most of the students were in the 6th-8th grades except for one class of 5th graders. Across the school district, the ethnicity of students enrolled (as of the 2021-2022 school year) was as follows: 62% White, 30% Hispanic, 3% Asian, 3% two or more races, 1% Black, 0.3% American Indian or Alaska Native, and 0.1% Hawaiian/Pacific Islander [77]. About half (49%) were female.



```
on start
    show icon
    set strip to NeoPixel at pin P12 with 5 leds as RGB (GRB format)
    strip set brightness 50
forever
    if sound intensity > 500 then
        clear screen
        strip set brightness 50
        strip show bar graph of sound intensity up to 2000
        show icon
    else
        show icon
        strip show color black
```

Figure 2. Screenshot of the MakeCode programming interface

2.1.3 Microphone

Our choice of microphone was influenced by several factors including audio quality, cost, power source, form-factor, and ease of use. We evaluated a range of candidates (e.g., MXL, Sony ICD PX370, ZOOM H1n, AudioTechnica-ATR, AudioTechnica-Omni, AudioTechnica-Cardioid, ProCon, Saramonic), we chose the Yeti Blue. This microphone has a user-configurable effective pickup pattern: omnidirectional, polar, XY stereo or cardioid, costing around $100USD. It is USB-powered, enabling use with an iPad without the need for an external mixer or phantom power.

2.1.4 Procedure

During each class, researchers placed microphones (Yeti Blue) at tables around the classroom. Groups of 2-4 consenting students were seated at each table. Depending on the lesson the students either worked as a team or as multiple dyads (during pair programming). The microphone was placed on the table roughly equidistant from all students, using the omnidirectional setting when recording 3 or more students, or the stereo setting when two students were sitting at either side of the microphoneThe microphone was connected to an iPAD via USB that hosted the recording software recording at 48kHz sampling rate. . We also collected iPAD video from a subset of students. Microphones were set up by a researcher who recorded field notes on different events (e.g., start of lesson, start of small group work, technical failures).

2.2 Data Treatment

2.2.1 Sample Selection.

We opted to select recordings with both audio and video to aid in ground-truth speaker diarization efforts (i.e., who is speaking). Of a total of 118 recordings, 79 met this criterion, of which we selected 49 recordings which contained small-group work, where at least one student in each group appeared in a minimum of 4 recordings.

From each video, five 5-minute candidate samples were selected from within the small-group work segment of the lesson, constrained to the middle of the segment such that the random sample included the midpoint of the small-group interval as beginning and end of the task tended to include less on-topic speech. A researcher then listened to each of the five random samples in turn. If the sample met the 20-word criterion, then it was selected for the sample. If it did not meet the criteria (*n* = 17), the next segment was listened to and so forth. If none of the 5 segments met the criteria, then the recording was excluded entirely. Through this process, we ended with 32 samples totaling 160 minutes of speech from 30 students. A majority (70%) of the students were in at least two recordings (Table 1).

Table 1. Sample summary

	M (SD)	Range
No. students per recording	2.6 (0.7)	2-4
No. recordings per student	2.7 (1.6)	1-5
No. utterances per recording	61.6 (29.6)	21-139
No. utterances per student	65.7 (47.2)	10-188
Wordcount per utterance	4.55 (4.03)	1-47

2.2.2 Transcription and annotation

Samples were transcribed in ELAN annotation software by trained transcribers, who recorded millisecond-resolution timestamps (based on the audio waveform) for utterance start and end times along with speaker identity. Where speaker identity was clear, but speech was too indistinct to transcribe, some or all of the utterance content was coded as "[inaudible]". This resulted in 2207 student utterances, of which 1970 contained at least 1 audible word (See Table 1).

Utterance-level audio segments were automatically transcribed by three cloud-based ASR services: Google Speech-to-text [26], Rev.ai [68], and IBM Watson [36]. We selected Google because it has been shown to work as well for children as adults [70] and in a recent review was shown to outperform similar services [20]. Watson has been used in multiple publications for ASR transcription of teacher talk [5, 37, 38] and as input to CPS linguistic models [63, 64]. Rev.ai was used as they claim equal or greater performance than Google [69]. We deemed these three ASRs sufficient for the present purposes of investigating patterns in and downstream influences of ASR errors and not to evaluate all available commercial ASR engines.

For Google, audio was first segmented using the human-segmented utterance-boundaries and individually submitted to the ASR. We used the video-optimized model as this was determined to outperform the default model in preliminary testing. For Rev.ai and Watson, all utterances from a given recording were concatenated before transcribing, as this theoretically allows the models to use prior language context to boost performance. The ASR result contains word-level timestamps which were used to split the full transcript back into the original utterances. We also tested using per-utterance transcripts for Watson and the Google streaming speech recognition API using the *single_utterance=True* option optimized for short utterances. Due to poorer performance than the main Watson and Google models, these were not analyzed further.

2.3 Measures

Before computing measures on the transcripts, all texts (human and ASR transcribed) were normalized to facilitate comparison. Non-word indicators used by the transcribers and ASR systems such as "[inaudible]", "[redacted]" and "%HESITATION" were stripped out. Numbers were spelled out if transcribed as digits. Leading and trailing punctuation was stripped from each word, and hyphens replaced by space. Finally, all words were transposed to lowercase.

2.3.1 Word Error Rate.

Using standard procedures [83], for each utterance, we used the Levenshtein algorithm at the word-level, which finds the minimum number of word substitution (S), insertion (I) and deletion (D) operations to align the reference (human transcript) to the hypothesis (ASR transcript). We used word error rate (WER) as a measure of transcription accuracy, which is given by: $WER = (S + D + I) / N_{\text{reference}}$ (number of words in the reference text). Proportion of insertion, substitution, and deletion errors were computed by dividing utterance-level error counts with the number of words in the human transcript. We also computed the number of words in the ASR transcripts along with a binary variable indicating whether the ASR returned any transcript at all.

2.3.2 Downstream NLP Measures

We focused on NLP tasks at the word, semantic, and discourse levels. In each case, we are interested in the error (distance) between the ASR (hypothesized) and human (reference) values.

2.3.2.1 BLEU Scores

The BLEU metric was developed to assess the performance of machine translation systems by comparing a gold standard translation to an output translation [59]. BLEU scores quantify sentence similarity based on modified n-gram precision, where scores vary from 0 (no match) to 1 (perfect match). This captures higher-order structure than WER: it is invariant to n-gram order, and encapsulates longer subsequences than WER which is defined only at the individual word level. We computed the BLEU score for unigrams, bigrams, trigrams, and quad-grams and computed an unweighted average of the four, which was reversed (i.e., 1-BLEU) to get the BLEU distance (or error)

2.3.2.2 Topic Word Analysis.

At the word level, we quantified students' uses of topic words that might be indicative of their cognitive engagement with the sensor immersion unit. The curriculum materials consist of storyboards, lesson plans, tutorials, etc., from which we generated a frequency dictionary of the named entities using the Named Entity Recognition algorithm from the Stanford CoreNLP toolkit [55]. Functional words were removed, resulting in 2,438 candidate words. Next, we used an existing Latent Dirichlet Allocation (LDA) topic model (created for an auxiliary purpose), which learns distinct topics from the document and returns the top 20 words that have the highest correlation with each of 20 topics. We computed the intersection of the 400 topic words and the 218 candidate named entities that occurred more than 20 times. This threshold ensured candidate words appeared at multiple points in the curriculum documents while keeping the list to a manageable size. This produced a set of 66 initial topic words. These topic words were then reviewed by curriculum experts who selected a subset of 33 topic words aligned to the following categories: science (e.g. *environmental),* coding (e.g. *function)*, and wiring (e.g. *sensors)*. For each utterance, we computed the number of topic words recognized by each ASR and the human transcript. To measure ASR fidelity specific to topic words, we compute *Topic Distance* as the absolute difference in utterance-level topic word counts between the human and ASR derived transcripts, with a lower bound of 0 and an undefined upper bound.

2.3.2.3 Semantic Distance

Beyond words themselves, we also evaluated the ASR transcripts using the semantic distance metric, which measures the similarity of a reference and a hypothesis transcript in a sentence-level embedding space (using a pre-trained language model to obtain the embeddings), and has been shown to be a better predictor of performance on downstream NLP tasks than traditional metrics such as WER [43]. Following the procedure outlined in [43], we first extracted utterance-level embeddings using the *sentence-transformers* Python library [67] and the '*all-distilroberta-v1*' model [50]. Then, we computed the cosine distances between the embeddings of each ASR (hypothesized) transcript and the reference human transcript. The cosine distance is defined as 1-cosine similarity (which ranges from -1 to 1), so it can take on values from 0 (identical) to 2 (dissimilar). To obtain a baseline value, we randomly shuffled the human transcripts within each 5-minute recording, then computed the semantic distance to each ASR transcript as described above. The average semantic distance over all ASRs was used as a baseline.

2.3.2.4 CPS Skill Classification

At the discourse level, we evaluated the utility of our ASR transcripts for a concrete NLP application: classifying collaborative problem solving (CPS) skills from student transcripts, which is one of the target applications noted in Section X. Specifically, we applied an existing classifier [63], which was trained to identify the following three CPS skills based on a validated CPS framework [80]: constructing shared knowledge; negotiation/coordination; maintaining team function, to our dataset. The classifier was a pre-trained BERT [16] model fine-tuned on a data set of 31,533 expert-coded student utterances (transcribed using the Watson ASR). Although the classifier was trained on a different dataset, it has been shown to be generalizable across domains [63], so we deemed it suitable for the present purposes. As such, we submitted both the human and ASR transcripts to the classifier, which outputs the predicted probabilities for the three CPS facets on each utterance. For each ASR, we computed the three-dimensional Euclidean distance between the ASR- and human- (reference) predicted probabilities as a measure of dissimilarity (CPS Distance). We also obtained a baseline shuffled value similar to the baseline Semantic Distance.

Table 2. Sample sentences and their corresponding ASR transcriptions. CPS codes: Const. = constructing shared knowledge; Neg. = negotiation/coordination; Maintain. = maintaining team function

Speaker	Human Transcript	Google	Watson	Rev	CPS Code
A	just start with the show number	start remove the show number	system started with the show numbers	start with the show number	Const.
B	oh	-	okay	okay	None
A	okay so you get rid of the show number	okc get rid of the sheriff	okay	okay so you get rid of the sharon remember	Maintain.
A	just drag it	stretch	dr don't	-	Maintain.
C	don't don't do that	don't don't do that	don't don't don't do that	don't do that	Maintain.
A	get rid of it	-	-	get rid of it	Maintain.
C	just okay	just	okay	it	Neg.
B	and now put this in this thing	i am for this and this thing	okay but this in this thing yeah	put this and this thing	Const.
A	yes	-	yeah	-	Neg.
C	no now you eat a taco	you know how you eat a taco	yeah are you talking	you need to talk	Maintain.
B	no do i put it in there	-	-	-	Const.
C	yeah	-	-	-	Neg.
A	yes	-	-	-	Neg.

2.4 Data Treatment

All measures (proportions of insertion, substitution and deletion errors; BLEU distance, Topic distance, semantic distance, CPS distance) were averaged per speaker per recording, resulting in 82 observations per ASR. This was done to obtain more reliable estimates due to the principle of aggregation. Because the distance metrics are only meaningful for utterances where the ASR returns a nonempty transcript, the averages for BLEU, Topic, Semantic and CPS distances were computed over nonempty transcripts only. To analyze the effects of ASR service and word errors on downstream measures, we used mixed effects linear regression models with speaker and recording identifier as random intercepts to account for the nested and repeated structure of the data with multiple speakers nested within recordings. Further, we used the r*obustlmm* package in R [45], which provides estimates that are robust to outliers and other contaminants in the data. We used estimated marginal means (*emmeans* package in R) for pairwise comparisons using false-discovery rate adjustments for multiple comparisons and Satterthwaite's degrees of freedom method. We used two-tailed tests with a $p < .05$ cutoff for significance.

3. RESULTS

3.1 ASR Errors

3.1.1 Patterns in Error Types

Table 3 provides descriptives on ASR performance measures averaged by student by recording. Immediately apparent is that the vast majority of ASR errors were deletion errors (67%) compared to substitution (17%) and insertion errors (6%; the sum of errors does not add up to 100% because of words correctly recognized). Indeed, when error rate was regressed on error type (three level categorical variable) and number of words in the human transcript (as a covariate), we found the following significant (*p*s < .001) pattern in the errors: Deletion > Substitution > Insertion (Table 3).

3.1.2 Comparing ASR Engines

We regressed each error type on ASR (a three-level categorical effect with Google as the reference group) and reference (human) transcript word count as a covariate. For deletion errors, we found Watson and Rev to be statistically equivalent and higher than Google suggesting the following significant (*p*s < .0001, FDR correction for 3 tests) pattern in the data: [Watson = Rev; $p = .61$] > Google. This pattern was largely flipped for substitution errors: Google > Watson > Rev; $p < .004$. For insertion errors, Google resulted in more insertion errors than Watson, but Rev was intermediate and not significantly different from either. Deletion errors ($p < .001$) were less likely as reference word count increased, but substitution ($p = .582$) and insertion ($p = .137$) errors were not.

Since insertion errors were rare, the tradeoff involved deletion and substitution errors (Figure 3) with Google providing fewest deletions but the most substitutions, the opposite for Rev, and Watson was intermediate. All things equal, the choice of ASR thus depends on obtaining as many transcriptions of speech as possible. Google provided a non-empty transcript for on average 61% of the cases, far exceeding the others (47% for Watson, 41% for Rev), and even among the utterances with nonempty ASR transcriptions, Google had a lower rate of deletions (0.29) than Rev (0.33) and Watson (0.44).

Table 3. Summary statistics of ASR results. M (SD) over utterances

	Google	Rev	Watson
N utterances	1970	1970	1970
N averaged	82	82	82
ASR metrics			
ASR wordcount	2.37 (1.36)	1.71 (1.46)	1.31 (0.95)
Nonempty ASR	0.61 (0.22)	0.41 (0.22)	0.47 (0.22)
Perfect ASR	0.05 (0.06)	0.04 (0.05)	0.02 (0.06)
Insertion rate	0.06 (0.09)	0.08 (0.16)	0.04 (0.08)
Substitution rate	0.21 (0.11)	0.12 (0.08)	0.19 (0.13)
Deletion rate	0.56 (0.18)	0.72 (0.17)	0.72 (0.17)
WER	0.84 (0.15)	0.91 (0.19)	0.95 (0.11)
Downstream NLP metrics			
Topic Distance	0.05 (0.10)	0.05 (0.09)	0.05 (0.07)
BLEU Distance	0.83 (0.11)	0.82 (0.15)	0.94 (0.06)
Semantic Distance	0.56 (0.14)	0.52 (0.15)	0.68 (0.09)
CPS Distance	0.29 (0.14)	0.29 (0.15)	0.33 (0.16)

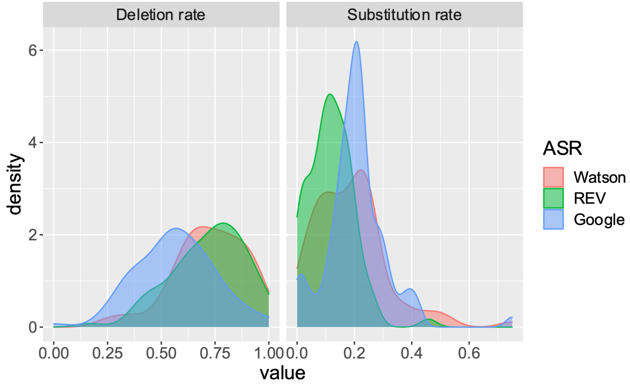


Figure 3. Density plots of deletion and substitution errors by ASR

3.1.3 Sources of Variance

We carried out a multilevel decomposition of variance [35] on each type of ASR error, at three levels: utterance, speaker and recording. Utterances are nested within speakers, and speakers within recordings. We computed the proportion of variance attributable to speaker and recording by decomposing the data into a linear sum of cluster-level averages and within-cluster deviations. The variance between-cluster and within-cluster sums to the total variance, under the assumption that errors at utterance, speaker, and recording are independent. We found that the majority of variance (between 91 and 98%) was at the utterance level for all error types and ASRs, with just 1-3% attributable to individual students and 1-5% to the specific recording (Table 4). This suggests that each ASR system had stable performance across recording contexts and individual differences in vocal parameters.

Table 4. Multilevel variance decomposition. Proportion of variance attributable to each hierarchical level.

ASR	Error type	Utterance	Student	Recording
Google	Insertion rate	0.981	0.009	0.010
Google	Substitution rate	0.980	0.010	0.010
Google	Deletion rate	0.943	0.032	0.025
Rev	Insertion rate	0.980	0.004	0.016
Rev	Substitution rate	0.971	0.009	0.020
Rev	Deletion rate	0.914	0.032	0.053
Watson	Insertion rate	0.976	0.009	0.015
Watson	Substitution rate	0.972	0.011	0.017
Watson	Deletion rate	0.950	0.025	0.025

3.2 Downstream NLP measures

The Spearman correlations between distance metrics for the four downstream tasks are shown in Table 5. The most highly correlated metrics were semantic distance and BLEU distance (r = .83), whereas the CPS distance was only moderately correlated (*rs* between .3 and .4) with these measures. Topic distance was not correlated with any other metric, which may be a result of topic words being so rare in the utterance (about 5% of words).

Table 5. Correlations between transcript distance metrics. *** p<0.001

	BLEU Distance	Topic Distance	Semantic Distance
Topic Distance	-0.118		
Semantic Distance	0.830***	-0.040	
CPS Distance	0.321***	0.076	0.402***

Figure 4 shows the distributions of CPS and semantic distances. The peak of the distribution was lower than the baseline (derived by computing the average distances between ASR and human-transcribed utterances after shuffling; see Methods) for all three ASRs and for both CPS and Semantic distances. In fact, 97.5% of semantic distances were less than the shuffled baseline, and 79% for CPS, indicating that a degree of higher-order meaning was generally extracted from the ASR transcripts.

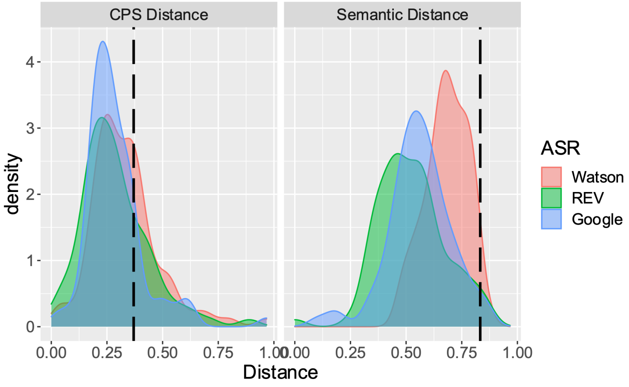


Figure 4. Density plots of CPS Distance and Semantic Distance by ASR. Dashed line shows the random baseline for shuffled utterances.

3.2.1 Comparing ASRs on Downstream NLP

We regressed each distance metric on ASR (a three-level categorical effect with Google as the reference group) and reference (human) transcript word count, with random intercept of student and recording. As indicated in Figure 5, the ASR services did not vary for Topic word distance (*p* = .929), but did for the other measures. Specifically, the pattern of significance (*ps* < .001) for BLEU and semantic distances was: Watson > Google > Rev. For CPS distance it was Watson > Rev, *p* = 0.03; Watson = Google, *p* = 0.16; Rev = Google, *p* = 0.46.

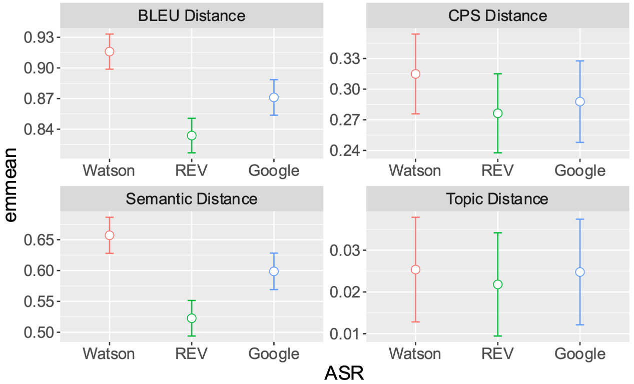


Figure 5. Estimated marginal means and 95% confidence intervals for NLP distance metrics for each of the ASR services

3.2.2 ASR errors on downstream NLP tasks

To test whether specific ASR errors impact downstream NLP metrics, we also fit a linear mixed-effects model to predict each distance metric from the rates of three ASR errors (Table 6). Whereas insertion errors did not significantly predict any of the outcomes, both substitution and deletion errors were negatively (*ps* < .001) associated with BLEU, semantic, and CPS distance measures, more so for the former. Specifically, a one standard deviation increase in each error type was associated with an approximately equivalent increase for BLEU and semantic distances, but only about a half a standard deviation increase for the CPS tasks. Error type was not associated with topic word distance, presumably due to a restriction of range with this measure.

Table 6. Mixed-effects model predicting distance metrics from ASR errors, showing standardized Beta values

	BLEU Distance		Topic Distance		Semantic Distance		CPS Distance	
Predictors	***std. Beta***	***p***	***std. Beta***	***p***	***std. Beta***	***p***	***std. Beta***	***p***
(Intercept)	0.07	<0.001	-0.27	0.020	0.05	<0.001	-0.05	0.189
Substitution rate	0.91	<0.001	-0.05	0.231	1.04	<0.001	0.46	<0.001
Deletion rate	1.08	<0.001	-0.05	0.256	1.12	<0.001	0.49	<0.001
Insertion rate	-0.03	0.345	0.01	0.676	0.03	0.446	-0.01	0.870

4. DISCUSSION

We investigated the feasibility of using commercially available ASRs to transcribe student discourse from a collaborative learning activity in a middle school classroom with an eye for downstream NLP tasks aimed to support student learning. In the remainder of this section, we discuss our main findings, applications, limitations, and areas for future work.

4.1 Main Findings

Overall, WER was very high (.84-.95) compared to performance on benchmark datasets, and even compared to WER from prior CL studies using classroom audio, such as in Pugh et al. 2021 who reported a WER of .78 using Watson, but with individual microphones in a more restricted in-class data collection setting compared to the current in-the-wild classroom context. At first blush, these high WERs suggest that it might be futile to expect meaningful ASR in noisy classroom environments without explicitly instrumenting the classroom for this purpose [1] or resorting to miking individual students with customized high-fidelity microphones [90]. However, an in-depth analysis of the pattern of errors suggests that there is hope: specifically, the ASRs had a large proportion of deletion errors and fewer substitution and almost no insertion error meaning that they tended towards high precision but low recall and are thereby feasible for applications that match this profile (as elaborated below).

Comparing the three ASR engines we examined, Google and Rev were biased towards more substitutions and deletions respectively, but also relevant is the proportion of utterances which did not get transcribed at all. Here, Google provided a clear advantage with nonempty results returned for 60% of utterances compared to less than 50% for the other two. Reassuringly, the variance in ASR errors was overwhelmingly from utterance-level differences, with very little attributable to recording or student. In addition, of 1970 utterances, only 477 (24%) returned no transcript from any ASR, raising the possibility of combining outputs from multiple ASRs.

We computed several distance metrics to capture ASR quality as reflected in downstream NLP measures, in each case computing the deviations between ASR-produced and human-transcript versions of each measure. With respect to the four measures, topic word usage was rare and there was very little variability in this measure so unsurprisingly there were no differences for it. Turning to the other measures, Watson was consistently outperformed by Google and Rev, which were equivalent on CPS distance. However, BLEU and semantic distance, which were strongly correlated, were best captured by Rev, despite Google having lower word-level error rates. Thus, Rev had a slight edge over Google for the downstream NLP tasks, but not sufficient to compensate for its higher deletion rate. Finally, as the NLP analyses got more abstract, ASR errors had less of an impact. The effect of substitutions and deletions on CPS distance (a discourse-based construct) was about half that of semantic and BLEU distances, (i.e. word- and semantic-level measures). Whereas this pattern is intuitively plausible it awaits replication with additional downstream NLP measures.

4.2 Applications

The ability to automatically capture and transcribe student speech during CL activities in the classroom opens the door for numerous applications. Fair and accurate ASR transcripts are the first step for automated interventions that aim to support CL in classrooms. One promising strand of research involves designing teacher-facing applications, such as teacher dashboards, which convey information about student collaborative talk to the teacher. The design space for such technologies is broad and relatively unexplored. While there is potential for abuses such as increased monitoring and evaluation of student talk, responsible innovations can also leverage student transcripts to celebrate students' contributions, build communities within classrooms and foster authentic collaboration motivated by student interest, not desire for positive reinforcement. For example, information gathered from CL discourse could be presented to a teacher offline (i.e., after class), illustrating any number of relevant details about the CL activity (e.g., what students talked about when on-task versus off-task, balance of speaking time, quality of collaboration). To demonstrate, we created an example dashboard visualization of the model-estimated occurrence of three CPS facets in student utterances (Figure 4) using both human- and Google- generated transcripts. As evident in the figure, model estimations are notably impacted by ASR error (i.e., in this group, the model underestimates the use of constructing shared knowledge by students A and C). Although model estimations will be imperfect, they can still provide valuable insights, and the impact of errors can be diminished by aggregating over longer time scales. These after-action reviews could greatly benefit teachers, giving them insight into how they might better support CL in their classroom. This includes designing new activities to better engage students, understanding which student groups may need additional support in future classes and what CPS skills students need help developing.

Similarly, these insights could be conveyed to the teacher online (i.e., during class) via a real-time teacher dashboard. Real-time feedback on CL groups could also enhance a teacher's ability for more effective classroom orchestration by providing them with novel insights into how groups are working together and what kinds of feedback and encouragement will help increase productive collaboration for students. Ultimately, the specifics of these teacher-facing applications, such as what information to present, when to present it (e.g., real-time, offline), how to display it (e.g., graphic representations, transcripts of speech) and at what level of granularity (e.g., individual students, CL groups, whole class) will require co-design, testing, and refinement with teachers.

In addition to teacher-facing applications, ASR systems could be used to create student-facing CL supports in the classroom. These technologies could take many forms, from real-time or after-action feedback that helps students develop CPS skills to a conversational agent which serves as a socio-collaborative 'partner', working together with student groups to enhance learning, equitable participation, and collaboration. Current approaches to support student collaboration (for example by prompting for the use of high-quality discourse called academically productive talk) have been shown to be successful in the context of text chat [85]. Further, after-action reviews to support CPS by providing feedback based on ASR/NLP models has demonstrated potential in the lab [64], but has yet to be tested in classrooms. Whether this is applied to student- or teacher-facing tools, fair and accurate ASR in classrooms has the potential to spotlight students' verbally-expressed ideas and contributions. This offloads the demand that is normally placed on written work and provides more multimodal dimensions for classroom feedback and support.

Whereas perfect ASR should not be a prerequisite for several applications (as argued in the Introduction), the patterns in ASR errors should be carefully considered in that the ASRs have high precision (relatively low substitution and insertion errors) but low recall (high levels of deletion errors). This suggests that these data are best suited for applications for which transcription of a sampling of utterances is sufficient, for example, assessments of constructs with high-base rates (e.g., CPS skills) rather than those focused on rare events. This high precision could be helpful in avoiding unwarranted interventions triggered by CL supports, as there should be a low rate of false alarms of discourse features detected based on the ASR results. Nevertheless, our findings suggest that a real-time conversational partner will likely be off the table until ASR deletion errors can be reduced. Nevertheless, robustness of NLP models to ASR errors can be improved by data augmentation approaches where models are trained on ASR hypotheses as well as human transcripts [58].

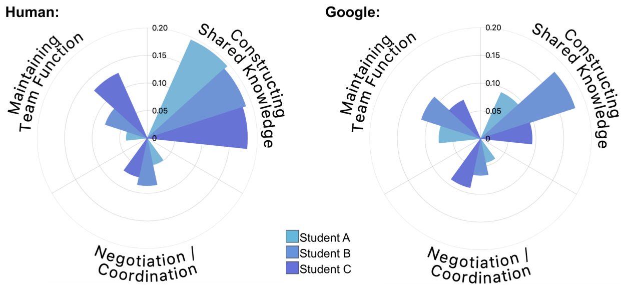


Figure 5. Predicted probabilities of CPS skills over a sample 5-minute recording (using the student group shown in Table 2) based on human (left) and Google ASR (right) transcripts

4.3 Limitations and Future Work

There were several limitations to this study. First, although we investigated an automated approach to transcribe student utterances, we did not incorporate automatic utterance segmentation in our pipeline. Rather, utterances were manually identified and segmented by a human observer before being processed by the ASR systems, which is consistent with prior work on comparing ASRs [5, 18]. This was done because the present focus was on speech-to-text transcription and not utterance-segmentation, so we opted for a gold-standard baseline for the latter to compare the various ASRs for the former. Further, utterance-segmentation is technically not needed as a separate step in an automated pipeline in that the entire five-minute audio segment could be submitted to the ASR engines for combined utterance segmentation and speech transcription, albeit less accurately than human segmentation. Indeed, longer context than single utterances are beneficial in modeling CL [64].

Another limitation is that we only tested out-of-the-box cloud-based ASR systems. One problem with this approach is that reliance on cloud-based services may be unrealistic in the near-term. In the US, nearly 28 million students did not have sufficient internet bandwidth for multimedia learning [97]. Similarly, we did not attempt to improve the performance of these out-of-the-box systems (e.g., by fine-tuning a custom ASR model on our data or providing a task-specific vocabulary) because the present goal was to compare these systems "as-is" since many researchers might not have the technical expertise needed to train customized models or fine-tune existing models. However, recent advances in deep-learning-based ASR mean pretrained models are widely available and a relatively small amount of data is needed for fine tuning [2, 74], which may provide better performance in this domain than standard cloud-based systems. To this point, we are currently developing a customized, locally hosted ASR system to improve upon the present results and address the limitations above.

One additional limitation is the lack of diversity in our sample. Whereas student-level demographic data was unavailable, district-level information suggests that 92% of the students were either White (62%) or Hispanic (30%). Racial disparities in ASR performance [44], as well as challenges with non-native English speakers [93] are well documented and may have disproportionately adverse effects on underrepresented groups when ASR is used for downstream applications. Thus, the lack of variability at the student level might be partly because our sample was non-representative. To create more fair ASR transcripts, non-native English speakers and students from non-dominant cultures should be oversampled to create representation, and thus accuracy, equal to students from dominant cultures. We also chose to include data from a single (although multi-lesson and multi-day) curriculum unit as implemented by one teacher with a small number of students. In sum, these factors reduce the generalizability of our findings to groups historically underrepresented in STEM. Our future work will aim to address these limitations by collecting classroom speech from racially and socioeconomically diverse populations, and examining ASR performance across different groups to identify sources of bias or nonequivalence.

4.4 Conclusion

Automated speech recognition in conjunction with natural language processing has the potential to unlock collaborative learning supports in the classroom. We recorded authentic small-group interactions in middle school STEM classrooms using inexpensive, commercially-available equipment, and analyzed the transcripts provided by several cloud providers. We show how different types of transcription errors influence downstream linguistic models, and find that the impact of ASR errors is smaller for the predictive accuracy of a CL model than for upstream measures capturing more literal aspects of speech content. Our results demonstrate the challenges of automating speech recognition in the classroom, but suggest the potential of using imperfect ASR to gain insights into collaborative discourse.

5. ACKNOWLEDGMENTS

This research was supported by the NSF National AI Institute for Student-AI Teaming (iSAT) (DRL 2019805). The opinions expressed are those of the authors and do not represent views of the funding agencies. We also thank Jon Cai (LDA topic models) and Robert Moulder (variance decomposition).

6. REFERENCES

[1] Ahuja, K., Kim, D., Xhakaj, F., Varga, V., Xie, A., Zhang, S., Townsend, J.E., Harrison, C., Ogan, A. and Agarwal, Y. 2019. EduSense: Practical Classroom Sensing at Scale. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*. 3, 3 (2019), 1–26. DOI:https://doi.org/10.1145/3351229.

[2] Baevski, A., Zhou, H., Mohamed, A. and Auli, M. 2020. wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations. *arXiv*. (2020).

[3] Bai, Y., Tejedor-García, C., Hubers, F., Cucchiarini, C. and Strik, H. 2021. An ASR-Based Tutor for Learning to Read: How to Optimize Feedback to First Graders. (2021), 58–69.

[4] Biddy, Q., Chakarov, A.G., Bush, J., Elliott, C.H., Jacobs, J., Recker, M., Sumner, T. and Penuel, W. 2021. Designing a Middle School Science Curriculum that Integrates Computational Thinking and Sensor Technology. *Proceedings of the 50th ACM Technical Symposium on Computer Science Education*. (Dec. 2021), 818–824. DOI:https://doi.org/10.1145/3287324.3287476.

[5] 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. *Proceedings of the 8th International Conference on Educational Data Mining 283* (2015), 23–33.

[6] Blanchard, N., D'Mello, S., Olney, A.M. and Nystrand, 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*. (2015), 282-288.

[7] Blanchard, N., Donnelly, P., Olney, A.M., Samei, B., Ward, B., Sun, X., Kelly, S., Nystrand, M. and D'Mello, S.K. 2016. Semi-Automatic Detection of Teacher Questions from Human-Transcripts of Audio in Live Classrooms. (2016).

[8] Bransford, J.D., Brown, A.L. and Cocking, R.R. 2000. How People Learn: Brain, Mind, Experience, and School.

[9] Caballero, D., Araya, R., Kronholm, H., Viiri, J., Mansikkaniemi, A., Lehesvuori, S., Virtanen, T., & Kurimo, M. (2017). Data Driven Approaches in Digital Education, 12th European Conference on Technology Enhanced Learning, EC-TEL 2017, Tallinn, Estonia, September 12–15, 2017, Proceedings. *Lecture Notes in Computer Science*, 541–544. https://doi.org/10.1007/978-3-319-66610-5_58

[10] Cazden, C.B. 1988. *Classroom discourse: The language of teaching and learning.* ERIC.

[11] Chakarov, A.G., Biddy, Q., Elliott, C.H. and Recker, M. 2021. The Data Sensor Hub (DaSH): A Physical Computing System to Support Middle School Inquiry Science Instruction. *Sensors (Basel, Switzerland)*. 21, 18 (2021), 6243. DOI:https://doi.org/10.3390/s21186243.

[12] Chang, X., Zhang, W., Qian, Y., Roux, J.L. and Watanabe, S. 2020. End-To-End Multi-Speaker Speech Recognition With Transformer. *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 00, (2020), 6134–6138. DOI:https://doi.org/10.1109/icassp40776.2020.9054029.

[13] Cole, R., Buchenroth-Martin, C., Weston, T., Devine, L., Myatt, J., Helding, B., Pradhan, S., McKeown, M., Messier, S., Borum, J. and Ward, W. 2018. One-on-one and small group conversations with an intelligent virtual science tutor. *Computer Speech & Language*. 50, (2018), 157–174. DOI:https://doi.org/10.1016/j.csl.2018.01.002.

[14] Cook, C., Olney, A.M., Kelly, S. and D'Mello, S.K. 2018. An Open Vocabulary Approach for Estimating Teacher Use of Authentic Questions in Classroom Discourse. *Proceedings of the 11th International Conference on Edu cational Data Mining*. (Jun. 2018), 116-126.

[15] Deng, A. 2021. *Fostering Literacy with Speech Recognition: A Pilot Study.*

[16] Devlin, J., Chang, M.-W., Lee, K. and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. (2019), 4171–4186. DOI:https://doi.org/10.18653/v1/n19-1423.

[17] Dillenbourg, P. (1999). What do you mean by collaborative learning? Chapter 1. In Dillenbourg (Ed.), *Collaborative-learning: Cognitive and Computational Approaches.* (Vol. 1, pp. 1–19). Oxford: Elsevier.

[18] 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. DOI:https://doi.org/10.1145/2818346.2830602.

[19] Feng, S., Kudina, O., Halpern, B.M. and Scharenborg, O. 2021. Quantifying Bias in Automatic Speech Recognition. *arXiv*. (2021).

[20] Filippidou, F. and Moussiades, L. 2020. A Benchmarking of IBM, Google and Wit Automatic Speech Recognition Systems. *Artificial Intelligence Applications and Innovations*. 583, (2020), 73–82. DOI:https://doi.org/10.1007/978-3-030-49161-1_7.

[21] Flor, M., Yoon, S.-Y., Hao, J., Liu, L. and Davier, A. von 2016. Automated classification of collaborative problem solving interactions in simulated science tasks. *Proceedings of the 11th Workshop on Innovative Use of NLP for Building Educational Applications*. (2016), 31–41. DOI:https://doi.org/10.18653/v1/w16-0504.

[22] Foltz, P.W., Laham, D. and Derr, M. 2003. Automated Speech Recognition for Modeling Team Performance. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*. 47, 4 (2003), 673–677. DOI:https://doi.org/10.1177/154193120304700402.

[23] Gamper, H., Emmanouilidou, D., Braun, S. and Tashev, I.J. 2020. Predicting Word Error Rate for Reverberant Speech. *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 00, (2020), 491–495. DOI:https://doi.org/10.1109/icassp40776.2020.9053025.

[24] German, S. 2019. Using the Anchoring Phenomenon Routine to introduce a science unit. *Science Scope*. 42, 5 (2019), 32–35.

[25] Gerosa, M., Giuliani, D., Narayanan, S. and Potamianos, A. 2009. A review of ASR technologies for children's speech.

Proceedings of the 2nd Workshop on Child, Computer and Interaction - WOCCI '09. (2009), 7. DOI:https://doi.org/10.1145/1640377.1640384.

[26] Google Cloud Speech-to-Text: *https://cloud.google.com/speech-to-text/*. Accessed: 2022-03-04.

[27] Graesser, A. C., Greiff, S., Stadler, M., & Shubeck, K. T. (2019). Collaboration in the 21st Century: The Theory, Assessment, and Teaching of Collaborative Problem Solving. *Computers in Human Behavior*, *104*, 106134. https://doi.org/10.1016/j.chb.2019.09.010

[28] Graesser, A.C., Person, N.K. and Magliano, J.P. 1995. Collaborative dialogue patterns in naturalistic one-to-one tutoring. *Applied Cognitive Psychology*. 9, 6 (1995). DOI:https://doi.org/10.1002/acp.2350090604.

[29] Griffin, P., Care, E. and McGaw, B. The changing role of education and schools. In P. Griffin, B. McGaw, & E. Care (Eds.), *Assessment and Teaching of 21st Century Skills* (pp. 1-16). Dordrecht, Germany: Springer http://dx.doi.org/10.1007/978-94-007-2324-5_2

[30] Hao, J., Chen, L., Flor, M., Liu, L. and Davier, A.A. von 2017. CPS-Rater: Automated Sequential Annotation for Conversations in Collaborative Problem-Solving Activities. *ETS Research Report Series*. 2017, 1 (2017), 1–9. DOI:https://doi.org/10.1002/ets2.12184.

[31] Hardy, L., Dixon, C. and Hsi, S. 2019. From Data Collectors to Data Producers: Shifting Students' Relationship to Data. *Journal of the Learning Sciences*. 29, 1 (2019), 1–23. DOI:https://doi.org/10.1080/10508406.2019.1678164.

[32] Henri, F. 1992. Computer Conferencing and Content Analysis. *Collaborative Learning Through Computer Conferencing*.

[33] Hmelo-Silver, C. E., & Barrows, H. S. (2008). Facilitating collaborative knowledge building. *Cognition and Instruction*, *26*(1), 48–94. https://doi.org/10.1080/07370000701798495

[34] Howard, C.S., Munro, K.J. and Plack, C.J. 2010. Listening effort at signal-to-noise ratios that are typical of the school classroom. *International Journal of Audiology*. 49, 12 (2010), 928–932. DOI:https://doi.org/10.3109/14992027.2010.520036.

[35] Husson, F., Josse, J., Narasimhan, B. and Robin, G. 2019. Imputation of Mixed Data With Multilevel Singular Value Decomposition. *Journal of Computational and Graphical Statistics*. 28, 3 (2019), 552–566. DOI:https://doi.org/10.1080/10618600.2019.1585261.

[36] IBM Watson: *https://www.ibm.com/watson/services/speech-to-text/*. Accessed: 2022-03-04.

[37] Jensen, E., Dale, M., Donnelly, P.J., Stone, C., Kelly, S., Godley, A. and D'Mello, S.K. 2020. Toward Automated Feedback on Teacher Discourse to Enhance Teacher Learning. *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. (2020), 1–13. DOI:https://doi.org/10.1145/3313831.3376418.

[38] Jensen, E., Pugh, S.L. and D'Mello, S.K. 2021. A Deep Transfer Learning Approach to Modeling Teacher Discourse in the Classroom. *LAK21: 11th International Learning Analytics and Knowledge Conference*. (2021), 302–312. DOI:https://doi.org/10.1145/3448139.3448168.

[39] Jeong, H., & Hmelo-Silver, C. E. (2016). Seven Affordances of Computer-Supported Collaborative Learning: How to Support Collaborative Learning? How Can Technologies Help? *Educational Psychologist*, *51*(2), 247–265. https://doi.org/10.1080/00461520.2016.1158654

[40] Jiang, D., Chen, Y. and Garg, A. 2018. A hybrid method for overlapping speech detection in classroom environment. *Computer Applications in Engineering Education*. 26, 1 (2018), 171–180. DOI:https://doi.org/10.1002/cae.21855.

[41] Johnson, R., Severance, S., Penuel, W.R. and Leary, H. 2016. Teachers, tasks, and tensions: lessons from a research–practice partnership. *Journal of Mathematics Teacher Education*. 19, 2–3 (2016), 169–185. DOI:https://doi.org/10.1007/s10857-015-9338-3.

[42] Kelly, S., Olney, A.M., Donnelly, P., Nystrand, M. and D'Mello, S.K. 2018. Automatically Measuring Question Authenticity in Real-World Classrooms. *Educational Researcher*. 47, 7 (2018), 451–464. DOI:https://doi.org/10.3102/0013189x18785613.

[43] Kim, S., Arora, A., Le, D., Yeh, C.-F., Fuegen, C., Kalinli, O. and Seltzer, M.L. 2021. Semantic Distance: A New Metric for ASR Performance Analysis Towards Spoken Language Understanding. *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH* (2021), 1977–1981.

[44] Koenecke, A., Nam, A., Lake, E., Nudell, J., Quartey, M., Mengesha, Z., Toups, C., Rickford, J.R., Jurafsky, D. and Goel, S. 2020. Racial disparities in automated speech recognition. *Proceedings of the National Academy of Sciences of the United States of America*. 117, 14 (2020), 7684–7689. DOI:https://doi.org/10.1073/pnas.1915768117.

[45] Koller, M. (2016). robustlmm : An R Package for Robust Estimation of Linear Mixed-Effects Models. *Journal of Statistical Software*, *75*(6), 1–24. https://doi.org/10.18637/jss.v075.i06

[46] Lee, S., Potamianos, A. and Narayanan, S. 1999. Acoustics of children's speech: Developmental changes of temporal and spectral parameters. *The Journal of the Acoustical Society of America*. 105, 3 (1999), 1455–1468. DOI:https://doi.org/10.1121/1.426686.

[47] Li, H., Ding, W. and Liu, Z. 2020. Identifying At-Risk K-12 Students in Multimodal Online Environments: A Machine Learning Approach. *arXiv*. (2020).

[48] Lileikyte, R., Irvin, D. and Hansen, J.H.L. 2020. Assessing Child Communication Engagement via Speech Recognition in Naturalistic Active Learning Spaces. *The Speaker and Language Recognition Workshop (Odyssey 2020)*. (2020), 396–401. DOI:https://doi.org/10.21437/odyssey.2020-56.

[49] Ling, H., Han, P., Qiu, J., Peng, L., Liu, D. and Luo, K. 2021. A Method of Speech Separation between Teachers and Students in Smart Classrooms Based on Speaker Diarization. *2021 13th International Conference on Education Technology and Computers*. (2021), 53–61. DOI:https://doi.org/10.1145/3498765.3498774.

[50] Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L. and Stoyanov, V. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv*. (2019).

[51] Loukina, A., Madnani, N., Klebanov, B.B., Misra, A., Angelov, G. and Todic, O. 2018. Evaluating on-device ASR on Field Recordings from an Interactive Reading Companion. *2018 IEEE Spoken Language Technology Workshop (SLT)*. 00, (2018), 964–970. DOI:https://doi.org/10.1109/slt.2018.8639603.

[52] Lugini, L., Olshefski, C., Singh, R., Litman, D. and Godley, A. 2020. Discussion Tracker: Supporting Teacher Learning about Students' Collaborative Argumentation in High School Classrooms. *Proceedings of the 28th International Conference on Computational Linguistics: System Demonstrations*. (2020), 53–58. DOI:https://doi.org/10.18653/v1/2020.coling-demos.10.

[53] Ma, Y., Wiggins, J.B., Celepkolu, M., Boyer, K.E., Lynch, C. and Wiebe, E. 2021. The Challenge of Noisy Classrooms: Speaker Detection During Elementary Students' Collaborative Dialogue. (2021), 268–281.

[54] MacNeilley, L.H., Nystrand, M., Gamoran, A., Kachur, R. and Prendergast, C. 1998. Opening Dialogue: Understanding the Dynamics of Language and Learning in the English Classroom. *Language*. 74, 2 (1998), 444. DOI:https://doi.org/10.2307/417942.

[55] Manning, C., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S. and McClosky, D. 2020. The Stanford CoreNLP Natural Language Processing Toolkit. *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*. 55–60. DOI:https://doi.org/10.3115/v1/p14-5010.

[56] Mutlu, B., Tscheligi, M., Weiss, A., Young, J.E., Kennedy, J., Lemaignan, S., Montassier, C., Lavalade, P., Irfan, B., Papadopoulos, F., Senft, E. and Belpaeme, T. 2017. Child Speech Recognition in Human-Robot Interaction. *Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction*. (2017), 82–90. DOI:https://doi.org/10.1145/2909824.3020229.

[57] National Teacher and Principal Survey: 2022. *https://nces.ed.gov/surveys/ntps/tables/ntps1718_fltable06_t1s.asp*.

[58] Nechaev, Y., Ruan, W. and Kiss, I. 2021. Towards NLU Model Robustness to ASR Errors at Scale. (2021).

[59] Papineni, K., Roukos, S., Ward, T. and Zhu, W.-J. 2002. BLEU: a method for automatic evaluation of machine translation. *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics - ACL '02*. (2002), 311–318. DOI:https://doi.org/10.3115/1073083.1073135.

[60] Pascual, R.M. 2020. Effectiveness of an Automated Reading Tutor Design for Filipino Speaking Children. *2020 IEEE 8th R10 Humanitarian Technology Conference (R10-HTC)*. 00, (2020), 1–5. DOI:https://doi.org/10.1109/r10-htc49770.2020.9357059.

[61] Potamianos, A. and Narayanan, S. 2003. Robust recognition of children's speech. *IEEE Transactions on Speech and Audio Processing*. 11, 6 (2003), 603–616. DOI:https://doi.org/10.1109/tsa.2003.818026.

[62] Praharaj, S., Scheffel, M., Drachsler, H. and Specht, M. 2019. Literature Review on Co-Located Collaboration Modeling Using Multimodal Learning AnalyticsCan We Go the Whole Nine Yards? *IEEE Transactions on Learning Technologies*. 14, 3 (2019), 367–385. DOI:https://doi.org/10.1109/tlt.2021.3097766.

[63] Pugh, S.L., Rao, A.R., Stewart, A.E.B. and D'Mello, S.K. 2022. Do Speech-Based Collaboration Analytics Generalize Across Task Contexts? *Learning and Knowledge 22*. (Apr. 2022), 208-218.

[64] Pugh, S.L., Subburaj, S.K., Rao, A.R., Stewart, A.E.B., Andrews-Todd, J. and D'Mello, S.K. 2021. Say What? Automatic Modeling of Collaborative Problem Solving Skills from Student Speech in the Wild. *Proceedings of The 14th International Conference on Educational Data Mining (EDM 2021*. (Mar. 2021), 55-67.

[65] Qian, Y., Weng, C., Chang, X., Wang, S. and Yu, D. 2018. Past review, current progress, and challenges ahead on the cocktail party problem. *Frontiers of Information Technology & Electronic Engineering*. 19, 1 (2018), 40–63. DOI:https://doi.org/10.1631/fitee.1700814.

[66] Reeder, K., Shapiro, J., Wakefield, J. and D'Silva, R. 2015. Speech Recognition Software Contributes to Reading Development for Young Learners of English. *International Journal of Computer-Assisted Language Learning and Teaching (IJCALLT)*. 5, 3 (2015), 60–74. DOI:https://doi.org/10.4018/ijcallt.2015070104.

[67] Reimers, N. and Gurevych, I. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. *arXiv*. (2019).

[68] Rev Speech-to-Text API: *https://www.rev.ai/*. Accessed: 2022-03-04.

[69] Rev vs Google ASR: *https://www.rev.com/blog/google-speech-recognition-api-vs-rev-ai-api*. Accessed: 2022-03-11.

[70] Rodrigues, A., Santos, R., Abreu, J., Beça, P., Almeida, P. and Fernandes, S. 2019. Analyzing the performance of ASR systems. *Proceedings of the XX International Conference on Human Computer Interaction*. (2019), 1–8. DOI:https://doi.org/10.1145/3335595.3335635.

[71] Roschelle, J., Dimitriadis, Y., and Hoppe, U. (2013). Classroom orchestration: Synthesis. *Computers & Education*, *69*, 523–526. https://doi.org/10.1016/j.compedu.2013.04.010

[72] Roschelle, J. and Teasley, S.D. 1995. Computer Supported Collaborative Learning. *Computer Supported Collaborative Learning*. 69–97.

[73] Rosé, C., Wang, Y.-C., Cui, Y., Arguello, J., Stegmann, K., Weinberger, A. and Fischer, F. 2008. Analyzing collaborative learning processes automatically: Exploiting the advances of computational linguistics in computer-supported collaborative learning. *International Journal of Computer-Supported Collaborative Learning*. 3, 3 (2008), 237–271. DOI:https://doi.org/10.1007/s11412-007-9034-0.

[74] Shivakumar, P.G. and Narayanan, S. 2022. End-to-end neural systems for automatic children speech recognition: An empirical study. *Computer Speech & Language*. 72, (2022), 101289. DOI:https://doi.org/10.1016/j.csl.2021.101289.

[75] Smith, L. and Macgregor, J.T. 1992. What is Collaborative Learning ? *Assessment*. 117, 5 (1992), 1-11.

[76] Song, Y., Lei, S., Hao, T., Lan, Z. and Ding, Y. 2021. Automatic Classification of Semantic Content of Classroom Dialogue. *Journal of Educational Computing Research*. 59, 3 (2021), 496–521. DOI:https://doi.org/10.1177/0735633120968554.

[77] St Vrain Demographics: *https://edx.cde.state.co.us/SchoolView/DataCenter/reports.jspx?Dis=0470&_afrLoop=3057061758625614&_afrWindowMode=0&tab=pro&_adf.ctrl-state=nxj1z2w8s_4*. Accessed: 2022-03-09.

[78] Stewart, A.E.B., Keirn, Z. and D’Mello, S.K. 2021. Multimodal modeling of collaborative problem-solving facets in triads. *User Modeling and User-Adapted Interaction*. 31, 4 (2021), 713–751. DOI:https://doi.org/10.1007/s11257-021-09290-y.

[79] Stewart, A.E.B., Vrzakova, H., Sun, C., Yonehiro, J., Stone, C.A., Duran, N.D., Shute, V. and D’Mello, S.K. 2019. I Say, You Say, We Say. *Proceedings of the ACM on Human-Computer Interaction*. 3, CSCW (Sep. 2019), 1–19. DOI:https://doi.org/10.1145/3359296.

[80] Sun, C., Shute, V.J., Stewart, A., Yonehiro, J., Duran, N. and D’Mello, S. 2020. Towards a generalized competency model of collaborative problem solving. *Computers & Education*. 143, (2020), 103672. DOI:https://doi.org/10.1016/j.compedu.2019.103672.

[81] Suresh, A., Jacobs, J., Lai, V., Tan, C., Ward, W., Martin, J.H. and Sumner, T. 2021. Using Transformers to Provide Teachers with Personalized Feedback on their Classroom Discourse: The TalkMoves Application. *arXiv*. (2021).

[82] Suresh, A., Sumner, T., Jacobs, J., Foland, B. and Ward, W. 2019. Automating Analysis and Feedback to Improve Mathematics Teachers’ Classroom Discourse. *Proceedings of the AAAI Conference on Artificial Intelligence*. 33, (2019), 9721–9728. DOI:https://doi.org/10.1609/aaai.v33i01.33019721.

[83] Szymański, P., Żelasko, P., Morzy, M., Szymczak, A., Żyła-Hoppe, M., Banaszczak, J., Augustyniak, L., Mizgajski, J. and Carmiel, Y. 2020. WER we are and WER we think we are. *arXiv*. (2020).

[84] Tao, Y., Mitsven, S.G., Perry, L.K., Messinger, D.S. and Shyu, M.-L. 2019. Audio-Based Group Detection for Classroom Dynamics Analysis. *2019 International Conference on Data Mining Workshops (ICDMW)*. 00, (2019), 855–862. DOI:https://doi.org/10.1109/icdmw.2019.00125.

[85] Tegos, S., Demetriadis, S. and Karakostas, A. 2015. Promoting academically productive talk with conversational agent interventions in collaborative learning settings. *Computers & Education*. 87, (2015), 309–325. DOI:https://doi.org/10.1016/j.compedu.2015.07.014.

[86] Terenzini, P.T., Cabrera, A.F., Colbeck, C.L., Parente, J.M. and Bjorklund, S.A. 2001. Collaborative Learning vs. Lecture/Discussion: Students’ Reported Learning Gains*. *Journal of Engineering Education*. 90, 1 (2001), 123–130. DOI:https://doi.org/10.1002/j.2168-9830.2001.tb00579.x.

[87] Tissenbaum, M. and Slotta, J.D. 2019. Developing a smart classroom infrastructure to support real-time student collaboration and inquiry: a 4-year design study. *Instructional Science*. 47, 4 (2019), 423–462. DOI:https://doi.org/10.1007/s11251-019-09486-1.

[88] Tissenbaum, M. and Slotta, J.D. 2015. Seamless Learning in the Age of Mobile Connectivity. *Seamless Learning in the Age of Mobile Connectivity*. 223–257.

[89] Vygotsky, L.S. 1978. *Mind and society: The Development of Higher Mental Processes*. Harvard University Press.

[90] Wang, Z., Miller, K. and Cortina, K. 2013. Using the LENA in Teacher Training: Promoting Student Involement through automated feedback. *Unterrichtswissenschaft*. 4, (2013), 290–305.

[91] 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. DOI:https://doi.org/10.1016/j.compedu.2014.05.010.

[92] Ward, W., Cole, R., Bolaños, D., Buchenroth-Martin, C., Svirsky, E., Vuuren, S.V., Weston, T., Zheng, J. and Becker, L. 2011. My science tutor: A conversational multimedia virtual tutor for elementary school science. *ACM Transactions on Speech and Language Processing (TSLP)*. 7, 4 (2011), 18. DOI:https://doi.org/10.1145/1998384.1998392.

[93] Wu, Y., Rough, D., Bleakley, A., Edwards, J., Cooney, O., Doyle, P.R., Clark, L. and Cowan, B.R. 2020. See What I’m Saying? Comparing Intelligent Personal Assistant Use for Native and Non-Native Language Speakers. *22nd International Conference on Human-Computer Interaction with Mobile Devices and Services*. (2020), 1–9. DOI:https://doi.org/10.1145/3379503.3403563.

[94] Xu, S., Ding, W. and Liu, Z. 2020. Automatic Dialogic Instruction Detection for K-12 Online One-on-One Classes. *Artificial Intelligence in Education*. 12164, (2020), 340–345. DOI:https://doi.org/10.1007/978-3-030-52240-7_62.

[95] Yasin, I., Liu, F., Drga, V., Demosthenous, A. and Meddis, R. 2018. Effect of auditory efferent time-constant duration on speech recognition in noise. *The Journal of the Acoustical Society of America*. 143, 2 (2018), EL112–EL115. DOI:https://doi.org/10.1121/1.5023502.

[96] Zylich, B. and Whitehill, J. 2020. Noise-Robust Key-Phrase Detectors for Automated Classroom Feedback. *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 00, (2020), 9215–9219. DOI:https://doi.org/10.1109/icassp40776.2020.9053173.

[97] *How people learn II: Learners, contexts, and cultures*. 2021. *REPORT ON SCHOOL CONNECTIVITY FUNDING YEAR 2021*. Connect K-12.