A Study of Automatic Speech Recognition in Noisy Classroom Environments for Automated Dialog Analysis

Nathaniel Blanchard[1], Michael Brady[1], Andrew M. Olney[2], Marci Glaus[3], Xiaoyi Sun[3], Martin Nystrand[3,] Borhan Samei[2], Sean Kelly[4], & Sidney D'Mello[1]

U. of Notre Dame[1], U. of Memphis[2], U. of Wisconsin-Madison[3], U. of Pittsburgh[4]

Abstract. The development of large-scale automatic classroom dialog analysis systems requires accurate speech-to-text translation. A variety of automatic speech recognition (ASR) engines were evaluated for this purpose. Recordings of teachers in noisy classrooms were used for testing. In comparing ASR results, Google Speech and Bing Speech were more accurate with word accuracy scores of 0.56 for Google and 0.52 for Bing compared to 0.41 for AT&T Watson, 0.08 for Microsoft, 0.14 for Sphinx with the HUB4 model, and 0.00 for Sphinx with the WSJ model. Further analysis revealed both Google and Bing engines were largely unaffected by speakers, speech class sessions, and speech characteristics. Bing results were validated across speakers in a laboratory study, and a method of improving Bing results is presented. Results provide a useful understanding of the capabilities of contemporary ASR engines in noisy classroom environments. Results also highlight a list of issues to be aware of when selecting an ASR engine for difficult speech recognition tasks.

Keywords: Google Speech, Bing Speech, Sphinx 4, Microsoft Speech, ASR engine evaluation

1 Introduction

Dialogic instruction, a form of classroom discourse focusing on the free exchange of ideas and open-ended discussion between teachers and students, has been linked to key constructs of learning such as student engagement [1] and deep comprehension [2]. Although classroom discussion is generally considered beneficial, actual use of appropriate dialogic instructional strategies in classrooms varies widely. Recent research in teacher education has demonstrated the importance of careful measurement and assessment of dialogic practices in promoting changes in teacher practice [3]. A long-term goal of the present research is to facilitate fast and efficient assessment of classroom discourse processes and outcomes for use in teacher professional development.

Motivation. Previously, large-scale efforts to improve classroom discourse have been conducted using complex, labor-intensive, and expensive excursions into classrooms. Nystrand and Gamoran [4, 5] studied thousands of students across hundreds of classroom observations of middle and high school English Language Arts classes. They

adfa, p. 1, 2011.

found positive effects on student learning from the overall dialogic quality of discourse. However, the sheer number of human coders required makes such studies cost prohibitive. The feasibility of such large-scale deployment has been stretched to its limits. With modern technology, a new approach consisting of the transcription of recorded in-class audio through automatic speech recognition (ASR), in combination with data mining and natural language understanding to automate the coding of classroom discourse, might finally be feasible.

We are addressing automated classroom dialogic analysis with CLASS 5. CLASS 5 is intended to automate dialogic instructional feedback through a large-scale implementation of Nystrand's coding scheme [4, 5], which focuses on the nature of questions involved in classroom discussion. Specifically, five properties of question events are coded: authenticity, uptake, level of evaluation, cognitive level, and question source. Nystrand et al. reported that among these variables, authenticity and uptake are the most important properties affecting student achievement [5, 6]. Previously, coders would begin by sitting in classrooms and recording when and what types of questions were asked followed by revision in the lab. CLASS 5 is intended to automate this task with an emphasis on recognizing different question events.

ASR is an important first step in recognizing question events from classroom audio because it enables the application of text-based machine learning techniques. However, speech recognition in noisy environments remains a challenging research problem, and most available ASR technologies are designed for desktop dictation in a quiet office environment. To determine the suitability of existing ASR technologies for CLASS 5, we analyze several out-of-the-box ASR solutions that do not require training on speakers and do not require any domain-specific knowledge. We focus on dialogic questions asked by teachers because they are highly correlated with student achievement [4]. This paper looks to identify which ASR systems are best suited for large-scale implementation of audio transcription in classrooms.

Related Work. Wang et. al. [7] experimented with real classroom audio in a way that could be adapted to provide feedback for teachers. They built classifiers to identify if 30-second segments of audio corresponded to discussion, lecture, or group work. Audio data was collected using LENA [8], which was adapted to report when either teachers are speaking, students are speaking, speech is overlapping, or there is silence; there was no attempt at ASR. Although Wang et. al. [7] reported success classifying classroom discourse at course-grained levels, their audio solution only provided information on *who was speaking*, while coding of question events requires knowing *what was said*.

Within the AIED community, much of the work on spoken-language technologies has focused on one-on-one interactions in intelligent tutoring systems (ITSs). For example, Litman and Silliman [9] developed ITSPOKE, a physics tutor that engages students in spoken discussion to correct errors and prompt student self-explanation. Mostow and Aist [10] built a reading tutor called Project LISTEN to improve oral reading and comprehension. Schultz et. al. [11] created a spoken conversational tutor architecture called SCoT. Ward et. al. [12] has developed a conversational ITS called My Science Tutor (MyST) for 3rd, 4th, and 5th grade students. Finally, Johnson and Valente

[13] created a spoken dialog tutor to teach language and cultural skills. Despite impressive advances in conversational tutoring, the focus of these systems has been one-on-one human-computer interactions with customized domain-specific desktop oriented ASR approaches. The question of whether these ASR solutions generalize in noisy classroom environments remains unanswered.

There have been some efforts to quantify contemporary ASR systems, albeit outside of classroom contexts. Morbini et al. [14] recently reviewed some of today's freely available ASR engines. They tested five ASR engines including Google Speech, Pocketsphinx, Apple, AT&T Watson, and Otosense-Kaldi. Tests were based on recordings obtained from six different spoken dialog systems, or systems where computers converse with humans in a variety of settings. These settings range from casual museum visitors speaking into a mounted directional microphone to trained demonstrators speaking into headset microphones. Their analyses focused on the strengths and weaknesses of the ASR engine's performance across different dialog systems. While their results provided a useful table of features associated with each engine, the authors concluded there was no single best ASR engine for all dialog systems. Their results did not address variable conditions that are often out of the developer's control, such as vocabulary domain. Furthermore, although the methods used to record audio were documented the quality and clarity of this audio were unreported. Thus, no inference can be drawn about which ASR engine would perform best for untested applications, such as transcribing naturalistic classroom discourse, thereby motivating the present study.

Contribution and Novelty. The present study provides, for the first time, a comparative evaluation and analysis of contemporary ASR engines for audio recordings from noisy classroom environments. The emphasis is on studying the accuracy of ASR engines on recordings of mic'ed teacher audio as they go about their normal classroom routines. We focus on teacher audio because dialogic instruction can be automatically coded using only teacher questions [15]. In addition to comparing transcription accuracy of five ASR systems, detailed analyses are performed on the two best-performing systems. The most effective ASR engine from the classroom study is validated in a follow-up laboratory study with more speaker variability. Although this work is done within the context of a specific research project (the development of CLASS 5), accurate ASR is important for many tasks. Taking a foundational look at what is possible for today's ASR systems in noisy classroom environments has implications for AIED researchers interested in developing other classroom-based spoken-language technologies or scaling up existing projects.

2 Classroom Recording Study

2.1 Method

Data Collection. Audio recordings were collected at a rural Wisconsin middle school during literature, language arts, and civics classes. The recordings were of three different teachers: two males – Speaker 1 and Speaker 2 – and one female – Speaker 3. The recordings span classes of about 45 minutes each on 9 separate days over a period of 3-

4 months. Due to the occasional missed session, classroom change, or technical problem, a total of 21 of these classroom recordings were available for analysis here. During each class session, teachers wore a Samson AirLine 77 ‘True Diversity’ UHF wireless headset unidirectional microphone that recorded their speech, with the headset hardware gain adjusted to maximum. Audio files were saved in 16 kHz, 16-bit mono .wav format. Teachers were recorded naturalistically as they taught their class as usual.

Two observers trained in Nystrand et. al.’s dialogic coding technique for audio annotation and classification [4, 16] were present in the classroom during recording. Observers marked teacher’s dialogic questions with start and stop times as the class progressed, and later reviewed the questions for accuracy. Audio of teacher questions was then extracted from the recordings and saved as individual .wav files by sectioning the audio using the observers’ labeled start and stop times. In total, there were 530 questions obtained from teacher speech. Table 1 presents information about the amount of time teachers spent asking questions (in seconds), mean verbosity (number of words in a question), mean duration (number of seconds taken to ask a question), mean speech rate (number of words per second), and maximum silence (longest pause in the middle of the speech). In general, Speaker 1 and Speaker 2 asked more questions than Speaker 3 and were more verbose. Speaker 3 had the slowest speech rate, while Speaker 2 tended to pause for the shortest amount of time when speaking.

Table 1. Means of question characteristics (standard deviations in parentheses)

Speaker	**N**	**Verbosity (words)**	**Question Duration (secs)**	**Speech Rate(words/sec)**	**Maximum Silence (secs)**
1	189	8.82 (6.06)	4.29 (2.29)	2.05 (0.79)	1.12 (0.52)
2	250	10.93 (6.82)	4.10 (2.30)	2.88 (1.25)	0.86 (0.60)
3	91	3.07 (2.48)	3.62 (1.60)	1.11 (0.89)	1.20 (0.64)

ASR Engines. We evaluated five ASR engines: Google Speech [17], Bing Speech [18], AT&T Watson [19], Microsoft Speech SDK 5.1, and two variants of Sphinx 4 [20]. 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). Google Speech includes twenty-seven languages and dialects, Bing Speech includes seven languages, and AT&T Watson includes nineteen languages and has recognition contexts. Sphinx 4 is a flexible ASR that allows developers to incorporate their own custom models; however, we limited our analysis to prebuilt acoustic models derived from the Wall Street Journal (Sphinx-WSJ), trained on people reading the WSJ, and the English Broadcast News Speech (Sphinx-HUB4), trained on speech from real broadcast news. Microsoft Speech, integrated with Windows since Vista, associates a speech profile with a user and adapts to that user’s speaking style and audio environment. We eliminated this adaptive bias by creating a new untrained speech profile for each recording date. We focus our efforts on these systems because they are freely available, except for Microsoft Speech (which requires a copy of Windows).

Evaluation Procedure. We processed all recorded questions through the ASR engines. We then compared the transcriptions that were output by the engines with observer-generated transcriptions. 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 coder's transcript and dividing the number of substitutions, insertions, and deletions required to transform the transcript into the hypothesis divided by the number of words in the transcript. SWO is the number of words that appear in both the computer-recognized speech and the human-recognized speech divided by the total number of words in the human-recognized speech. WAcc preserves word order while SWO ignores it. WAcc is bounded on $(-\infty, 1]$ while SWO is bounded on $[0, 1]$. For both metrics higher numbers indicate better performance.

2.2 Results

Overall ASR Accuracy Rates. Table 2 presents the mean WAcc and SWO by ASR. Here, the cloud-based ASR engines Google Speech and Bing Speech clearly outperformed the other engines. Google Speech performed 7.69% better than Bing Speech when word order was considered (WAcc metric), but Bing performed 3.33% better than Google when word order was ignored (SWO metric). Bing and Google WAcc was, respectively, 26.8% and 36.6%, higher than AT&T Watson. The Sphinx HUB4 model did show improvements over the WSJ model, but overall HUB4 accuracy was lower than Bing and Google, with a performance similar to Microsoft.

Given their superior performance on the two key metrics, we focus subsequent analyses on Google and Bing. Because WAcc was strongly correlated with SWO for both Bing (Pearson's r = 0.792) and Google (r = 0.908), we focus on WAcc.

Table 2. Mean accuracy by ASR (standard deviations in parentheses)

ASR	WAcc	SWO
Google Speech	0.56 (0.35)	0.60 (0.31)
Bing Speech	0.52 (0.41)	0.62 (0.31)
AT&T Watson	0.41 (0.48)	0.53 (0.31)
Sphinx (HUB4)	0.14 (0.61)	0.32 (0.30)
Microsoft	0.08 (0.70)	0.33 (0.31)
Sphinx (WSJ)	0.00 (0.67)	0.27 (0.27)

Error Types. On average, for Bing, 40% (SD = 36%) of the errors were substitutions (ASR substituted one word for another), 30% (SD = 36%) were deletions (ASR missed words), and 15% (SD = 26%) were insertions (ASR inserted words). For Google, 44% (SD = 36%) of errors were substitutions, 35% (SD = 36%) were deletions, and 8% (SD = 20%) were insertions. Thus, there were modest differences across ASR engines for substitution and deletion errors, and larger differences for insertion errors.

WAcc by Individual Speaker. Small differences were found between speakers. The mean difference in average WAcc across pairs of speakers (i.e., average of Speaker 1 vs. Speaker 2, Speaker 1 vs. Speaker 3, and Speaker 2 vs. Speaker 3) were quite small – 0.12 for Bing and 0.06 for Google. This suggests that these ASRs were mostly unaffected by speaker variability, at least with respect to the teachers in our sample

WAcc by Class Session. To quantify the consistency of the ASRs across class sessions, we conducted a decomposition of variance in error rates within and between class observations [21]. For Bing, 3.8% of the variance in error rates lies between class sessions; the vast majority of the variance in error rates was within-observations, across utterances. For Google, a similarly small percentage of variance lies between observations, only about 3.3% in these data. Thus, automatic speech recognition is largely invariant to the differences in instructional topic, etc. occurring in these data.

WAcc by Speech Characteristics. We investigated the relationship between WAcc and the four speech characteristics listed in Table 1. Models that regressed WAcc on these four speech characteristics (using a stepwise feature selection method) explained 2.6% of the variance for Google and 4.2% of the variance for Bing. The negligible variance explained in these models indicates Google and Bing were mostly immune to variation in speech characteristics.

Confidence of ASR Hypotheses. Bing provides confidence estimates with its output, thereby affording an additional analysis of Bing. We note mean WAcc scores of 0.00, 0.35, 0.48, 0.65 for confidence levels of: rejected (N = 4), low (N = 59), medium (N = 284), and high (N = 183), respectively. Removing the rejections and the low confidence questions resulted in a mean WAcc of 0.55 for the remaining 467 utterances, which reflects a small improvement over the overall WAcc of 0.52 reported in Table 1. The results were not more notable because confidence estimation itself was imperfect.

Comparing Google and Bing. Table 3 provides a cross tabulation for questions recognized perfectly (WAcc = 1), completely incorrectly (WAcc <= 0), and in between (0 < WAcc < 1) across both ASRs. Bing and Google completely failed and succeeded for the same 28 and 30 questions, respectively. Interestingly, Google perfectly transcribed 10 of the 78 questions that Bing completely failed to recognize, while Bing perfectly transcribed 16 of the 73 questions that Google completely failed to recognize. In general, Bing and Google's WAcc scores were only modestly correlated (Pearson's r = 0.306), which suggests that there may be advantages to combining them.

Table 3. Cross tabulation of Bing and Google WAcc

		Google		
		WAcc <= 0	*0 < WAcc < 1*	*WAcc = 1*
Bing	*WAcc <= 0*	28	40	10
	0 < WAcc <1	29	325	29
	WAcc = 1	16	23	30

Qualitative Analysis of Complete Failures. We performed a qualitative analysis on the questions for which both ASRs completely failed (WAcc <= 0) by listening to each audio file and noting potential causes of errors (See Table 4). The most common failure involved the teacher questioning a student by calling his or her name (e.g., “Marty?”) – both ASRs were equally susceptible to this issue. Another common failure, more so for Google than for Bing, occurred when the teacher was quizzing a student on specific vocabulary words that were either not in the ASR dictionaries or were rare enough to be unrecognizable without context (e.g., cacophony, despot). Bing often failed when audio was not perfectly segmented (e.g. the segmented audio file began in the middle of loud student speech), an inevitable byproduct of collecting audio in a noisy environment. Both ASRs experienced failure when questions were only one word, which typically occurred when students were quizzed on vocabulary. Google faltered when teachers rushed through questions. Bing experienced 19 complete failures when the teacher began a question but paused for a long interval before continuing with the question. Both recognizers failed when teachers asked implicit questions, such as where the teacher began a statement and paused for the student to complete the utterance (“speaker one is…”). The recognizers struggled when teachers over-enunciated syllables, which occurred when presenting unfamiliar vocabulary to students.

Table 4. Failure Analysis (Number of errors by category and ASR)

Error	Bing	Google	Error	Bing	Google
Student name	26	25	Rushed	5	18
Vocabulary	11	21	Long Pause	19	0
Imperfect Segmentation	23	12	Implicit question	4	6
One word	12	13	Over enunciate	6	3

WAcc Improvements by Eliminating Pauses for Bing. We identified 113 instances with imperfect WAcc likely attributed to a long teacher pause, which negatively affected Bing but not Google, as indicated by the analysis of complete failures (see Table 4). To mitigate these failures, pauses were automatically identified and removed, and the resulting modified audio was rerun through Bing. Eliminating the silences raised mean WAcc from 0.30 (SD = 0.41) to 0.34 (SD = 0.58), eliminated all rejections, and raised the overall WAcc for Bing from 0.52 (SD = 40.1) to 0.53 (0.46) and SWO from 0.62 (SD = 0.31) to 0.65 (SD = 0.30). Removing the remaining low confidence instances increased Bing’s WAcc to 0.58 (SD = 0.43) and Bing’s SWO increased to 0.71 (SD = 0.25). Thus, Bing’s WAcc was higher than Google’s overall WAcc.

3 Laboratory Study on Reliability Across Speakers Using Bing

Due to the logistics of data collection, the original classroom study only involved 3 speakers. We therefore conducted a laboratory study with 28 speakers to test the reliability of Bing across a larger number of speakers. We focused on Bing instead of Google because it has an easier to use application programming interface (API – details of which are not discussed here), it provides confidence scores, it resulted in WAcc performance equal to Google (after eliminating pauses and low confidence scores), and exhibits a SWO well above Google.

Method. 13 male and 15 female for a total of 28 English-speaking undergraduate students were recruited. These participants were instructed to play the part of a teacher leading a discussion in a classroom. Participants read the teacher's lines from transcripts of classroom speech displayed on a computer screen. The students' portions of the scripts were pre-recorded by an actor and automatically played in response to the participant's speech. Participants proceeded through three scripts constructed from transcripts of three separate teachers. Script order was balanced across participants with a Latin Square. Participant speech was recorded using the same headset microphone as was used to record teachers in classrooms.

Results. In total, 3057 recordings of dialog turns were obtained and submitted to Bing for recognition. Utterance-level mean WAcc was 0.60 (SD = 0.32), considerably higher than the previously reported Bing WAcc of 0.52 (SD 0.41), presumably because of the controlled laboratory environment. We performed a decomposition of variance analysis to quantify the variation in word error rates for Bing across participants. We found a small but non-trivial proportion of variance at the speaker level (ICC=.096). The standard deviation of the word error rates across speakers was 0.10 (about the grand mean of 0.61). The estimated confidence interval for the ICC suggests that as much as 15% of the variance in word error rates is a function of speaker-level speech attributes.

4 Discussion

We tested five implementations of ASR engines on 530 spontaneous spoken dialogic questions from 3 different teachers recorded with a headset microphone in noisy classrooms, and conducted a follow-up laboratory study. Google Speech and Bing Speech largely outperformed AT&T Watson, Microsoft Speech, and two implementations of Sphinx. A summary of results yields seven key insights:

1. Google ASR performed slightly better than Bing when word order was considered, but Bing performed slightly better than Google when word order was disregarded. The WAcc of Google and Bing was only moderately correlated, indicating different strengths and weaknesses, thereby raising the possibility of combining the two.
2. The majority of Bing and Google errors were substitution errors. Furthermore, deletion errors were more frequent than insertion errors.
3. Differences in speakers and sessions had little impact on WAcc. This conclusion was further validated for Bing using 28 English speakers in a laboratory study.
4. Bing was susceptible to failure when speech had long pauses. We corrected this by removing long pauses. Doing so, along with removing low confidence results, resulted in Bing having a higher WAcc than Google.
5. Speech characteristics (i.e. speech duration, speech verbosity, speech length, and maximum silence) were found to have very small effects on ASR accuracy. This indicates Google Speech and Bing Speech were largely unaffected by these factors.
6. Bing provides useful confidence scores that are relatively representative of the accuracy of the hypothesis, along with multiple alternative hypotheses about what was spoken. This is a major advantage of using Bing over Google.

7. Both Bing and Google completely failed or perfectly succeeded for a roughly 14% of utterances, and these instances of complete failure and success did not always overlap. Combining these two recognizers to avoid these failures may be strategic.

Limitations. Our results suggest that Bing and Google were the best out-of-the-box ASR engines for automatically transcribing teacher speech in noisy classrooms with a specific emphasis on questions. However, other applications may have different ASR needs. The engines we selected for evaluation were limited to free engines requiring no training or optimizing in any way. Furthermore, our study focused only on dialogic questions. We note the possibility that the strict use of questions in our study may somehow have influenced our results (though we have no reason to believe this is the case). Since we were able to collect recordings from only three teachers from one school, there is a chance that our results will not be corroborated across regions or speakers. However, considering the versatility of the ASRs thus far, a significant change in our results in response to a larger data set is not anticipated, but this awaits empirical verification. Finally, we did not test all possible ASR engines and we acknowledge that some ASR engines could perform better than the engines we tested.

Concluding Remarks. Our results give us some confidence that ASR technologies have matured to the point that they can be useful for the automatic transcriptions of speech from classrooms and other noisy environments. To be clear, these technologies are still far from perfect. However, the goal is not to obtain perfect transcription of speech, but to obtain a reasonable representation of spoken dialog to serve as input to language processing techniques that should be uninfluenced by ambiguities in speech recognition (see [22] for further discussion). Furthermore, our analysis of where these systems succeed and fail on recordings from the classroom environments should benefit researchers who have been working with educational dialog in on-one-one settings, but who are interested in testing spoken-language technologies in the classroom.

5 Acknowledgements

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.

6 References

1. Kelly, S.: Classroom discourse and the distribution of student engagement. Soc. Psychol. Educ. 10, 331–352 (2007).
2. Sweigart, W.: Classroom Talk, Knowledge Development, and Writing. Res. Teach. Engl. 25, 469–496 (1991).
3. Juzwik, M.M., Borsheim-Black, C., Caughlan, S., Heintz, A.: Inspiring Dialogue: Talking to Learn in the English Classroom. Teachers College Press (2013).
4. Nystrand, M., Gamoran, A., Kachur, R., Prendergast, C.: Opening dialogue. Teach. Coll. Columbia Univ. N. Y. Lond. (1997).

5. Gamoran, A., Kelly, S.: Tracking, instruction, and unequal literacy in secondary school English. Stab. Change Am. Educ. Struct. Process Outcomes. 109–126 (2003).
6. Nystrand, M., Gamoran, A.: The big picture: Language and learning in hundreds of English lessons. Open. Dialogue. 30–74 (1997).
7. Wang, Z., Pan, X., Miller, K.F., Cortina, K.S.: Automatic classification of activities in classroom discourse. Comput. Educ. 78, 115–123 (2014).
8. Ford, M., Baer, C.T., Xu, D., Yapanel, U., Gray, S.: The LENA Language Environment Analysis System. Technical Report LTR-03-2. Boulder, CO: LENA Foundation (2008).
9. Litman, D.J., Silliman, S.: ITSPOKE: An intelligent tutoring spoken dialogue system. Demonstration Papers at HLT-NAACL 2004. pp. 5–8. Association for Computational Linguistics (2004).
10. Mostow, J., Aist, G: Evaluating tutors that listen: An overview of Project LISTEN. (2001).
11. Schultz, K., Bratt, E.O., Clark, B., Peters, S., Pon-Barry, H., Treeratpituk, P.: A scalable, reusable spoken conversational tutor: Scot. Proceedings of the AIED 2003 Workshop on Tutorial Dialogue Systems: With a View toward the Classroom. pp. 367–377 (2003).
12. Ward, W., Cole, R., Bolaños, D., Buchenroth-Martin, C., Svirsky, E., Vuuren, S.V., Weston, T., Zheng, J., Becker, L.: My science tutor: A conversational multimedia virtual tutor for elementary school science. ACM Trans. Speech Lang. Process. TSLP. 7, 18 (2011).
13. Johnson, W.L., Valente, A.: Tactical Language and Culture Training Systems: using AI to teach foreign languages and cultures. AI Mag. 30, 72 (2009).
14. Morbini, F., Audhkhasi, K., Sagae, K., Artstein, R., Can, D., Georgiou, P., Narayanan, S., Leuski, A., Traum, D.: Which ASR should I choose for my dialogue system? Proceedings of the SIGDIAL 2013 Conference. pp. 394–403. , Metz, France (2013).
15. Samei, B, Olney, A, Kelly, S, Nystrand, M., D'Mello, S, Blanchard, N, Sun, X, Glaus, M, Graesser, A: Domain independent assessment of dialogic properties of classroom discourse. Proceedings of the 7th International Conference on Educational Data Mining. pp. 233–236. Stamper, J., Pardos, Z., Mavrikis, M., McLaren, B.M., London, England, UK (2014).
16. Nystrand, M., Wu, L.L., Gamoran, A., Zeiser, S., Long, D.A.: Questions in time: Investigating the structure and dynamics of unfolding classroom discourse. Discourse Process. 35, 135–198 (2003).
17. Schalkwyk, J., Beeferman, D., Beaufays, F., Byrne, B., Chelba, C., Cohen, M., Kamvar, M., Strope, B.: "Your Word is my Command": Google Search by Voice: A Case Study. Advances in Speech Recognition. pp. 61–90. Springer (2010).
18. Microsoft: The Bing Speech Recognition Control. (2014). http://www.bing.com/dev/en-us/speech. Accessed 14 Jan 2015
19. Goffin, V., Allauzen, C., Bocchieri, E., Hakkani-Tür, D., Ljolje, A., Parthasarathy, S., Rahim, M.G., Riccardi, G., Saraclar, M.: The AT&T WATSON Speech Recognizer. ICASSP (1). pp. 1033–1036 (2005).
20. Walker, W., Lamere, P., Kwok, P., Raj, B., Singh, R., Gouvea, E., Wolf, P., Woelfel, J.: Sphinx-4: A flexible open source framework for speech recognition. (2004).
21. Kelly, S., Majerus, R.: School-to-school variation in disciplined inquiry. Urban Educ. 0042085911413151 (2011).
22. D'Mello, S.K., Graesser, A., King, B.: Toward Spoken Human–Computer Tutorial Dialogues. Human–Computer Interact. 25, 289–323 (2010).