INTERSPEECH 2016
September 8–12, 2016, San Francisco, USA

The SRI Speech-Based Collaborative Learning Corpus

Colleen Richey[1], Cynthia D'Angelo[2], Nonye Alozie[2], Harry Bratt[1], Elizabeth Shriberg[1]

[1] SRI International Speech Technology and Research (STAR) Laboratory, United States
[2] SRI International Center for Technology in Learning (CTL), United States

{colleen.richey, cynthia.dangelo, maggie.alozie, harry.bratt, elizabeth.shriberg} @ sri.com

Abstract

We introduce the SRI speech-based collaborative learning corpus, a novel collection designed for the investigation and measurement of how students collaborate together in small groups. This is a multi-speaker corpus containing high-quality audio recordings of middle school students working in groups of three to solve mathematical problems. Each student was recorded via a head-mounted noise-cancelling microphone. Each group was also recorded via a stereo microphone placed nearby. A total of 80 sessions were collected with the participation of 134 students. The average duration of a session was 20 minutes. All students spoke English; for some students, English was a second language. Sessions have been annotated with time stamps to indicate which mathematical problem the students were solving and which student was speaking. Sessions have also been hand annotated with common indicators of collaboration for each speaker (e.g., inviting others to contribute, planning) and the overall collaboration quality for each problem. The corpus will be useful to education researchers interested in collaborative learning and to speech researchers interested in children's speech, speech analytics, and speech diarization. The corpus, both audio and annotation, will be made available to researchers.

Index Terms: speech corpus, automatic speech recognition, children's speech, collaborative learning, STEM education.

1. Introduction

The SRI Speech-Based Collaborative Learning Corpus is being collected as part of a project investigating the utility of a speech-based learning analytics approach to collaborative learning. Collaboration is a core teaching and learning process that students must master as they progress through school and their careers [1]. It has been investigated for several decades and a robust theory documenting key features of collaborative learning exists [2], [3]. Face-to-face collaborative and cooperative learning have demonstrated to be beneficial for students' learning [4]. However, collaboration is a process that is difficult to manage and assess in a typical classroom setting. Research in this area could increase knowledge of how humans and automatic speech recognition can judge the quality of student collaboration based on features of student speech. Such judgments are important to teachers' implementation of collaborative learning in classrooms and they make collaborative learning more quantifiable.

Assessing collaboration includes many facets. We focus on the speech component and the corpus is collected with that in mind. Our goal is determining whether detectable patterns exist in student speech that correlate with collaborative learning indicators and that provide a means of assessing collaboration quality. To that end, the corpus contains manual annotations (1) marking indicators of collaboration and (2) assessing the overall collaboration quality of the interaction.

We report on phase 1 of the speech corpus: speech collected from two small groups of students working simultaneously. Phase 1 data collection is complete and annotation is underway. Phase 2 will contain audio recordings of full classrooms (25-30 students) working simultaneously in small groups.

2. Data collection

The first phase of data collection is complete. The collection contains audio recordings from visits to 21 middle schools and the participation of 134 students. The corpus also contains associated software logs. Video recordings were collected for use in annotation. The audio recordings and software logs will be made available to researchers after the conclusion of the project in 2017.

2.1. Collaboration task

The participants in this data collection were middle school students (grades 6-8) from the San Francisco Bay Area. Participation took place after school in a mathematics classroom. They worked in groups of three on sets of short mathematics problems that require collaboration to solve. The mathematics problems were developed under an earlier project and have been extensively tested [5]. They were a collaborative variation of the cloze task (fill in the blank) [6], in which each student was assigned one blank and each problem required the students to work together and talk to each other to coordinate their three answers. Figure 1 has a screen shot showing one of the cloze tasks.

The collaborative mathematics problems were delivered on iPads with a custom-built software application. The software logs recorded the timestamps for the beginning and end of each problem, every answer choice each student made, every solution the group collectively agreed on, and whether that solution was correct.

Most students participated in two sessions with different group configurations. In session 1, students entered their answers via the touch screen; in session 2, students entered their answers via their own controller. The average length of a single session was 20.47 minutes.

http://dx.doi.org/10.21437/Interspeech.2016-1541

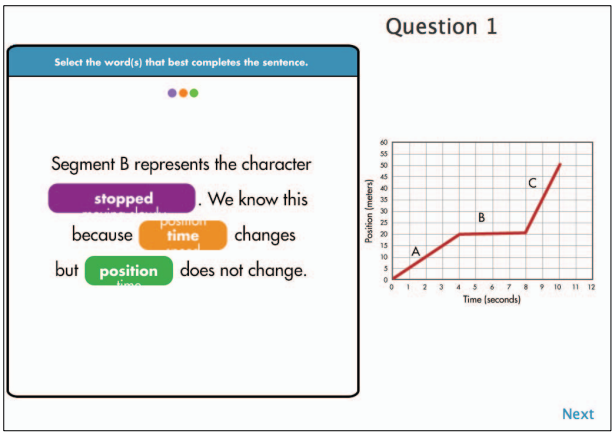

Figure 1: *Example screen shot*

2.2. Recording setup

Two groups of three students participated at the same time. Each group sat at separate tables that were situated as far apart as the classroom configuration allowed. Students sat in a row (from left to right: student 1, student 2, student 3) in front of an iPad activity station. Each student wore a head-mounted noise-cancelling microphone (Audio-Technica PRO 8HEx). Students 1 and 2 wore the headset so that microphone pointed away from the group in order to cancel as much audio from the other two students as possible. A ZOOM H6 portable digital recorder was used to record audio from these three microphones plus its built-in stereo microphone. All four audio channels were digitized at 48 kHz with 24-bit PCM encoding using a shared clock (and were therefore sample synchronous).

Each group was also video recorded from behind to capture the iPad screen and most hand gestures. The video was used for annotation purposes only and will not be shared.

Because the collection took place after school, background noise was limited. The recordings do contain noise from the other participating group, students outside the classroom, and intercom announcements.

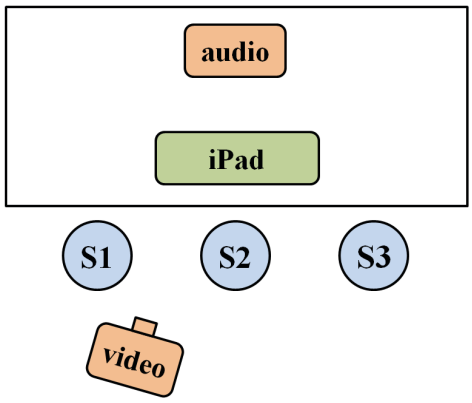

Figure 2: *Schematic drawing of recording setup*

3. Data segmentation

The audio (three mono and one stereo sample synchronous channels) has been aligned with the start and end times that each group worked on each mathematical problem or "item". This alignment was done initially by using the timestamps saved in the software logs and then manually based on the audio and video recordings.

The corpus contains timestamps for the speech regions from each participant based on the output of a speech activity detection (SAD) system [7], [8]. The spoken interactions contain much overlapping speech, so the SAD system was run separately on the audio from each student's microphone. The SAD system calculated a spectral variability score for every 10 ms of audio. Based on a small sample, we determined a spectral variability threshold and used that threshold to decide if speech was contained in each 10 ms audio window. Assuming minimum durations for the regions of speech and non-speech, we then concatenated these 10ms windows to form regions of speech from a single speaker. Each speech region roughly corresponds to a spoken utterance from a single student, though utterances with long pauses may be broken up into two or more regions.

In addition to automatically detecting the speech regions, we manually annotated the speech regions for a small subset of the sessions. We marked the start and end times for each region of speech from each speaker. The utterances were broken up if they contained long pauses. These manual timestamps can be used to obtain more accurate speech duration measures and to tune an automatic SAD system. All audio segmentation, both automatic and manual, will be made available to the researchers after the conclusion of the project in 2017.

4. Data annotation

All audio recordings made during active work on the mathematics problems are being manually annotated by education researchers for (1) indicators of collaboration and (2) overall collaboration quality. These annotations will enable researchers to investigate the correlation between speech features and certain verbal behaviors associated with collaboration (positively and negatively) and overall collaboration quality.

All audio has been manually annotated for overall quality of collaboration. At this paper's publication time, 94% of the audio has been annotated or indicators of collaboration. Annotations will be made available to researchers after the conclusion of the project in 2017.

4.1. Annotation procedure

To establish inter-rater reliability on the human annotations, the annotators participated in multiple training sessions over approximately 30 days for the collaboration quality (Q) codes and over approximately 60 days for the indicators of collaboration (I) codes. The training sessions served as a space for developing a shared meaning of the codes and more detailed criteria for each code, in order to establish reliability and to refine the coding scheme. The annotators watched videos of the collaboration and had access to the waveforms of the separate audio channels. The annotations were applied in a custom application that could show the video synced up with the audio channel data. After the training period, the annotators were assigned videos to code independently. The majority of the videos were coded by a single annotator and about 10% were coded by all team members for periodic

reliability checks. The annotation teams also met regularly to discuss coding issues.

4.2. Indicators of collaboration

Work on face-to-face collaboration [9] and online collaboration [10] lists behaviors that are "supportive of collaboration." These behaviors include giving and receiving help and assistance, exchanging resources and information, and explaining or elaborating information. We built off of these and other extant collaboration coding schemes to iteratively construct a coding scheme that would be appropriate for the type of task and group setting that we were using. The collaboration indicator (I) codes are organized into three categories: (1) regulative/logistical codes; (2) interaction-based codes; and (3) learning-related (cognitive) codes. Some codes were given a hypothetical valence as to whether they were expected to positively or negatively influence the overall collaboration quality of the group. The I codes are listed in Table 1.

Table 1: *Indicators of collaboration quality.*

Regulative/Logistical	Planning Monitoring progress Verbalizing thinking Reading problem aloud Communicating that thinking
Interaction	Turn sharing Being ignored Acknowledging
Learning-Related	Inviting others to contribute Cognitive frustration Asking a question Giving away an answer Explaining Agreeing Disagreeing Expressing lack of understanding

Annotators assigned I codes to the individual audio channels. For each I code, the annotators assigned a start time, an end time, and a label. Figure 3 shows a schematized example of I code annotation.

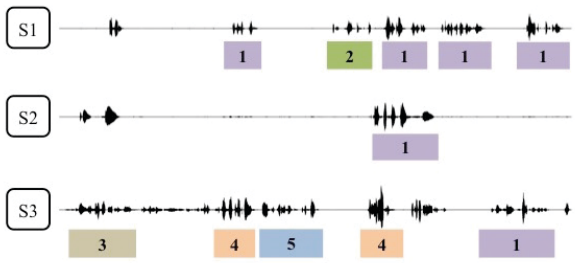

Figure 3: *Example of I codes aligned with audio from each student. 1=Verbalizing thinking; 2=Agreeing;3= Communicating that thinking; 4=Inviting others to contribute; 5=Planning*

4.3. Overall collaboration quality

The collaboration quality (Q) codes represent the degree to which the three students collectively were engaging in good collaboration. Importantly, from the perspective of the expert human coders, they depend not on how much each student talked but on whether and how much each student was engaged intellectually in the group problem solving. In descending order of collaboration quality, the four Q codes are: (1) good collaboration; (2) out in the cold; (3) follow the leader; and (4) not collaborating. Additionally, if students did not have an opportunity to collaborate (e.g., they were waiting for technical help) or if the window was too short to assess, then the coding window was marked as not applicable (N/A) for a Q code.

- Good Collaboration: All three students are working together and intellectually contributing to problem solving.
- Out in the Cold: Two students are working together, but the third is either not contributing or is being ignored.
- Follow the Leader: One student is taking the intellectual lead on solving the problem and is not bringing in others.
- Not Collaborating: No students are actively contributing to solving the problem (either off-task or independently working).

A team of five annotators (different annotators than for the I codes) made coding decisions at two levels: the item level (which varied in length depending on how long it took the group to solve the problem) and a fixed 30-second window within each item. All windows and items received a primary Q code, for the prominent quality of that time segment. When more than one mode of collaboration occurred in a segment, the annotators could optionally apply a secondary Q code for the other mode.

For the Q codes, the annotators reached more than 85% agreement (Cohen's kappa of 0.61) at the 30-second chunk level before working on their own. Following the training period, each annotator independently viewed and coded an assigned number of videos independently. To maintain consistency, 15% of the coded videos were checked by a third coder. Every two weeks during the coding period, the human coders reconvened to discuss coding consistency.

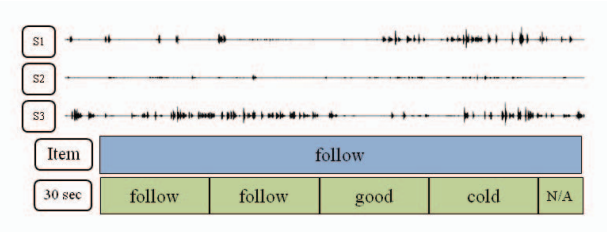

Figure 4: *Example of Q codes aligned with the audio*

5. Corpus statistics

5.1. Audio data

The data collection contains audio recordings from 80 sessions of groups of three students interacting. A total of 134 students participated with an approximately equal number for each gender – 68 females and 66 males. Half of the students were in

the 6th grade (67 students) and the remainder were in 7th grade (35 students) and 8th grade (32 students). Students included both native speakers of English and and English language learners. Most students participated in two sessions each. However, some school visits had more than six students participate, and some students had to sit out one session. Of the students, 106 participated in both sessions.

The average duration of a session was 20.4 minutes and the groups worked on an average of 10.9 items per session. The average duration of an item was 104.5 seconds. Based on the output of the automatic SAD system, the average amount of speech per student per session was 4.6 minutes. A large range existed in the time that different students spoke. The shortest amount of speech from a student in a single session was 3.6 seconds, while the most was 12.4 minutes. This range was also reflected in the number of regions of speech from each student in a single session. Based on SAD output, the students averaged 308.4 speech regions per session. The smallest number of speech regions for a student in a single session was 14 and the largest number was 754. Table 2 summarizes the numbers.

Table 2: *Amount of speech per speaker per session.*

Measure	Average	Minimum	Maximum
Duration of speech	4.6 min	3.6 sec	12.4 min
Number of speech regions	308.4	14	754

Table 3: *Distribution of I codes.*

I Code	Count	I Code	Count
Verbalizing thinking	9605	Explaining	349
Planning	3440	Turn sharing	286
Communicating that thinking	2578	Expressing lack of understanding	276
Agreeing	1920	Monitoring progress	266
Reading problem aloud	1249	Giving away an answer	181
Acknowledging	1239	Cognitive frustration	67
Disagreeing	1052	Inviting others to contribute	30
Asking a question	728	Being ignored	16

5.2. Annotation data

At time of publication almost all sessions have been annotated with I codes (75 out of 80 sessions). All sessions from phase 1 of the data collection will be annotated by the end of the project. So far, more than 23K I codes have been assigned, with an average of 310.5 I codes per session. Table 3 shows the distribution of I codes in descending order of frequency.

All 80 session have been annotated with Q codes. A relatively balanced distribution of the four Q codes exists, indicating that the data contain a range of collaboration quality among the groups of students. Table 4 shows the distribution of Q codes at the item level and at the 30-second window level.

Table 4: *Distribution of Q codes.*

Q Code	Item-Level Count	Window-Level Count
Good collaboration	352	1009
Out in the cold	224	804
Follow the leader	177	622
Not collaborating	118	519

6. Summary and future directions

In this paper, we introduce a new and novel speech corpus: The SRI Speech-Based Collaborative Learning Corpus. The corpus contains high-quality multi-channel sample-synchronous audio recordings of middle school students interacting while solving collaborative mathematical problems. The corpus contains automatic and manual segmentation of the audio into speech regions as well as careful annotation of indicators of collaboration and overall collaboration quality.

Phase 1 of the data collection has been completed and at time of publication almost all the audio has been annotated for overall collaboration quality and for indicators of collaboration. Research on predicting collaboration quality from non-lexical speech features is ongoing [11] [12]. Future work will investigate the correlation between (1) indicators of collaboration and overall collaboration quality and (2) indicators of collaboration with non-lexical speech features.

Phase 2 of data collection has been completed and annotation is underway. It is a more complicated collection that involves dividing a full classroom of students into small groups and recording all groups working simultaneously. The audio recordings contain more realistic noise levels, such as those expected during collaborative learning in the classroom. These recordings can be used to further assess the utility of speech-based methods of assessment in the classroom.

This corpus was designed specifically for researching speech patterns during collaborative learning and determining whether these features can provide a means to assess collaboration quality. The corpus will be of general interest to both education and speech researchers. The corpus, (audio, segmentation, and annotation) will be shared with others for research purposes.

7. Acknowledgements

We would like to acknowledge the contributions of Mingyu Feng, Jeremy Fritts, Reina Fujii, Amy Hafter, Diana Jang, Erik Kellner, Tiffany Leones, Daisy Rustein, Tina Stanford, and Jeremy Roschelle.This material is based upon work supported by the National Science Foundation under Grant No. DRL-1432606. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

8. References

[1] National Research Council, *Assessing 21st Century Skills: Summary of a Workshop*, Washington, D.C.: The National Academies Press, 2011.

[2] E. G. Cohen, "Restructuring the classroom: Conditions for productive small groups," *Rev. of Educ. Res.*, vol. 64, no. 1, pp. 1-35, 1994.

[3] P. A. Kirschner, and G. Erkens, Toward a framework for CSCL research, *Educational Psychologist*, *48*(1), 1-8, 2013.

[4] D. W. Johnson and R. T. Johnson, Learning together and alone: Overview and meta-analysis, *Asia Pacific Journal of Education, 22*(1), 95-105, 2002.

[5] J. Roschelle, N. Shechtman, D. Tatar, S. Hegedus, B. Hopkins, S. Empson, J. Knudsen and L. Gallagher, Integration of technology, curriculum, and professional development for advancing middle school mathematics: Three large-scale studies, *American Educational Research Journal, 47(4), 833-878*, 2010.

[6] W. L. Taylor, "Cloze procedure": A new tool for measuring readability, *Journalism Quarterly, 30*, 415-433, 1953.

[7] P. K. Ghosh, P. Kumar, A. Tsiartas and S. Narayanan, "Robust voice activity detection using long-term signal variability," *Audio, Speech, and Lang. Proc., IEEE Trans.*, vol. 19, no. 3, pp. 600-613, 2011.

[8] A. Tsiartas, T. Chaspari, N. Katsamanis, P. K. Ghosh, M. Li, M. Van Segbroeck, A. Potamianos and S. Narayanan, "Multi-band long-term signal variability features for robust voice activity detection," *Interspeech 2013*, pp. 718-722, 2013.

[9] D. W. Johnson and R. T. Johnson, "Cooperation and the use of technology," In D. H. Jonassen (Ed.), *Handbook of Research for Educational Communications and Technology* (pp. 1017-1044). New York: Simon and Schuster Macmillan, 1996.

[10] D. D. Curtis and M. J. Lawson, "Exploring collaborative online learning," *Journal of Asynchronous Learning Networks*, 5(1), 21-34, 2001.

[11] J. Smith, H. Bratt, C. Richey, N. Bassiou, E. Shriberg, A. Tsiartas, C. D'Angelo, and N. Alozie, "Spoken interaction modeling for automatic assessment of collaborative learning," in Speech Prosody, 2016.

[12] N. Bassiou, A. Tsiartas, J. Smith, H. Bratt, C. Richey, E. Shriberg, C. D'Angelo, and N. Alozie, "Privacy-Preserving Speech Analytics for Automatic Assessment of Student Collaboration," in *INTERSPEECH 2016 — 17th Annual Conference of the International Speech Communication Association, Proceedings*, San Francisco, California, USA, September 8-12, 2016.