Collaboration Detection that Preserves Privacy of Students' Speech

Sree Aurovindh Viswanathan(✉) and Kurt VanLehn(✉)

Arizona State University, Tempe, AZ 85281, USA
{sviswal0, kurt.vanlehn}@asu.edu

Abstract. Collaboration is a 21st Century skill as well as an effective method for learning, so detection of collaboration is important for both assessment and instruction. Speech-based collaboration detection can be quite accurate but collecting the speech of students in classrooms can raise privacy issues. An alternative is to send only whether or not the student is speaking. That is, the speech signal is processed at the microphone by a voice activity detector before being transmitted to the collaboration detector. Because the transmitted signal is binary (1 = speaking, 0 = silence), this method mitigates privacy issues. However, it may harm the accuracy of collaboration detection. To find out how much harm is done, this study compared the relative effectiveness of collaboration detectors based either on the binary signal or high-quality audio. Pairs of students were asked to work together on solving complex math problems. Three qualitative levels of interactivity was distinguished: Interaction, Cooperation and Other. Human coders used richer data (several audio and video streams) to choose the code for each episode. Machine learning was used to induce a detector to assign a code for every episode based on the features. The binary-based collaboration detectors delivered only slightly less accuracy than collaboration detectors based on the high quality audio signal.

Keywords: Collaborative learning · Machine learning · Learning analytics

1 Introduction

Collaboration is a 21^{st} century skill as well as an effective method for learning [1]. However, learning to collaborate is not straightforward. Students may require feedback to develop collaboration skills [2]. For scaling up feedback and assessment of collaboration, automated methods for collaboration detection are required. Fortunately, when students interact via speech, collaboration can be differentiated from non-collaboration using current technology, as reviewed below.

However, monitoring collaboration in spoken conversations between students raises concerns about privacy. Although teachers are always entitled to hear the speech of their students, giving third parties access to student conversations may raise significant privacy concerns. An alternate approach would be to process the audio signal at the microphone by using a voice activity detector (also called speech activity detector), which converts the raw audio signal into a binary signal (1 = Speech, 0 = Silence). Once the audio signal is converted to a binary signal, it can be transmitted or stored for

S. Isotani et al. (Eds.): AIED 2019, LNAI 11625, pp. 507–517, 2019.
https://doi.org/10.1007/978-3-030-23204-7_42

collaboration analysis. Because the binary signal is incomprehensible, privacy is preserved. However, the loss of information may prevent effective classification of collaboration.

This paper compares the relative effectiveness of measuring collaboration based on either a high quality audio signal or its binary version. Since the total amount of information transmitted by high quality signal is many orders of magnitude greater than the binary version, we expected a large difference in classifier performance. The high quality audio signal was collected by headset microphones connected to tablets being used by small groups of students who were solving problems in a laboratory setting. Only low level analysis was performed on both signals. Spoken words were not used as part of the analysis. For analysis of high quality speech data, low level features such as pitch, shimmer and linear spectral features were used. For the binary signals, time series features such as absolute energy, approximate entropy and symmetry were used. All features were extracted by algorithms and no human coders were involved.

In order to create and evaluate collaboration detectors, the judgments of human coders were used as the 'gold standard' classification of the group's interactions. The coders had both high quality audio and several videos to aid their judgment. Collaboration detectors were then machine-learned from the human judgments. Their accuracies were measured using 10 fold cross validation.

2 Prior Work on Speech-Based Collaboration Detection

Many systems have explored automated analyses of interaction among group members [3]. Instead of speech, most such systems input typed text from students collaborating via forums, chat or email. Of the projects that used speech-based collaboration detectors [4–11], only 3 measured the accuracy of the classification. These 3 projects are the most similar to our project, so they will be reviewed here.

Just as we did for our high quality speech classifier, Gweon et al. [6, 13] used machine-learned classifiers based on low-level speech features. Although the amount of speech and silence were included as features, the temporal pattern of speech and silence were not considered in their analysis. Secondly, whereas collaboration was the focal code, Gweon et al. two projects chose different non-collaboration codes. This choice may impact accuracy, so we measured the accuracy of classifiers trained with different combinations of non-collaboration codes.

Just as we did for our binary-signal classifier, Martinez Maldonado et al. [9] used a voice activity detector to convert speech into binary. However, they used counting and proportions in their classifiers; they did not consider temporal patterns. Later, this group did consider temporal patterns [4, 8]. They used differential sequence mining to find temporal patterns of speech and silence that would reliably split groups into high and low collaborators. However they did not convert their findings into a collaboration detector and measure its accuracy.

Bassiou et al. [11] conducted studies that are quite similar to ours. They found that collaboration detectors induced from a binary signal were more accurate than collaboration detectors induced from low-level acoustic features. They also found that accuracy could be further improved by combining the two types of features. Their studies differed from ours in several ways. Most importantly, they used a coarser coding scheme for collaboration. The scheme merely indicated how many students in the group were participating actively in problem solving.

3 Data Collection

This section describes the context in which the data were gathered.

3.1 Task: Collaborative Writing

The subjects collaboratively solved a problem ("Boomerangs") that was developed by the Mathematics Assessment Project and appears on their site [14]. Like prior work on spoken collaboration detection, it is a math problem. However, unlike the prior work, students are required to write paragraph-long explanations (See Fig. 1). They were given solutions to an optimization problem done by 4 hypothetical students and were asked 3 questions about each solution. Thus, this task is actually a collaborative writing task. It clearly has a different cadence than most mathematical problem solving. In particular, there can be significant periods of time when one student is writing out an explanation developed by one or both students. Because collaborative writing is required in many tasks from outside mathematics, it is important to investigate the accuracy of speech-based collaboration detection while student are thus engaged.

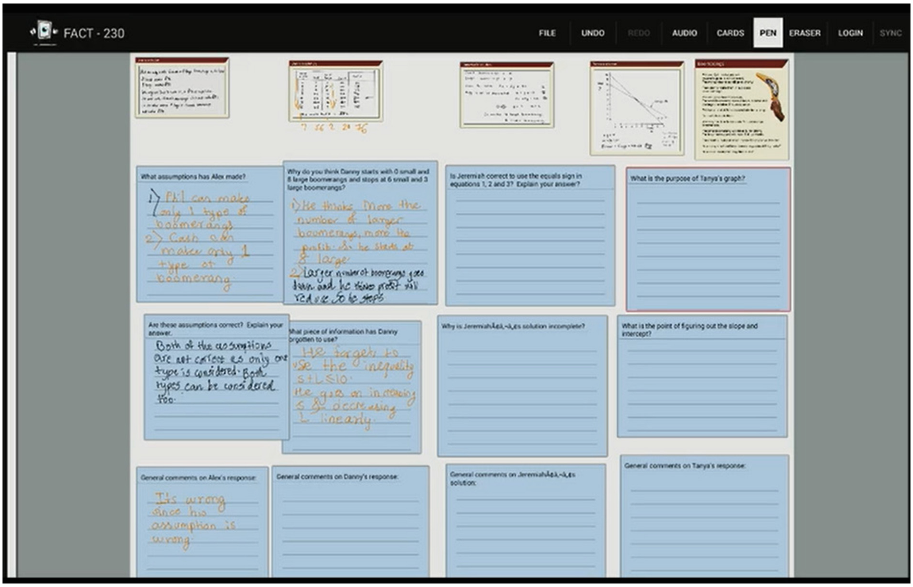
3.2 Technology, Participants and Duration

Students worked together in pairs. Each student had their own tablet, a Samsung Galaxy Note 10.1. This tablet had active digitizer technology which allowed students to write easily and legibly on the tablet screen using a stylus. These tablets are connected to the Server via a Wireless network. The software used by participants is called FACT [22]. The FACT user interface mimicked a large poster on which students can write and draw content. Anything written by one group member was immediately visible to the other. Both members of the group could scroll and zoom independently of each other thus allowing them to focus on different parts of the poster.

The study was conducted in a laboratory setting. The participants were 38 graduate and undergraduate students from our university who were paid for their time. Prior to doing the collaborative writing task, students solved the optimization problem themselves in order to become familiar with it. All students were able to solve it easily. The overall task, including both problem solving and collaborative writing, took around 45 to 55 min to complete.

3.3 Raw Data Collection

The recording setup generated input streams from two unidirectional headset microphones, one omnidirectional microphone, two tablet screen and two Web cameras. A desktop screen recorder was used to combine these input signals for easier annotation by the human coders.



What assumptions has Alex made?

1) Ph:1 can make only 1 type of boomerangs
2) Cath can makes only 1 type of boomerang

Why do you think Danny starts with 0 small and 8 large boomerangs and stops at 6 small and 3 large boomerangs?

1) He thinks More the number of larger boomerangs, more the profit. So he starts at 8 large
2) Larger number of boomerangs goes down and he thinks profit will reduce. So he stops

Is Jeremiah correct to use the equals sign in equations 1, 2 and 3? Explain your answer?

What is the purpose of Tanya's graph?

Are these assumptions correct? Explain your answer.

Both of the assumptions are not correct as only one type is considered. Both types can be considered too.

What piece of information has Danny forgotten to use?

He forgets to use the inequality S+L≤10. He goes on increasing S & decreasing L linearly.

Why is JeremiahÂ¢â‚¬â„¢s solution incomplete?

What is the point of figuring out the slope and intercept?

General comments on Alex's response:

It's wrong since his assumption is wrong.

General comments on Danny's response:

General comments on JeremiahÂ¢â‚¬â„¢s solution:

General comments on Tanya's response:

Fig. 1. Snapshot of students working on shared workspace

3.4 Coding Categories

This section describes the collaboration codes assigned by human coders. Because the coders could see all the input streams, whereas the collaboration detectors could only hear the audio from the two headset mics, the human coders' judgements were used as the "gold standard" against which the collaboration detectors were judged.

Several coding schemes for human judgement of collaboration have been devised [1, 11, 15–17]. The simplest merely count the number of members of the group that are actively participating. Thus, if both members of our groups were speaking or writing, then the group would be coded as collaborating.

However, even if all members of a group are participating, they may be interacting to varying degrees. On one end of this scale, they have split up the task and each member is working individually on a different subtask; this end of the scale is often called *cooperation*. On the other extreme, the members are essentially pursuing one line of reasoning or argumentation with all members contributing to it. That is, each member's utterance or action refers to and builds upon the prior contributions of other members. This end of the scale is often called *co-construction, transactivity* or just *collaboration*. The degree of interaction among group members manifests itself in different ways, so most coding schemes are multi-dimensional, where each dimension alerts coders to one way in which the cooperation/collaboration distinction can be observed.

When a collaboration detector is used in a classroom, it should help the teacher make a binary decision - whether to visit a group or not. Similarly, when it is used in a tutoring system, it should help the system make a binary decision – to intervene or not.

Different teachers and systems might have different concerns, but our sense of the literature is that most use cases can be covered if the collaboration detector outputs these three classifications:

1. *Interactive:* Students are working on the same part of the poster, and they contributions build upon each other. This is often called co-constructive or transactive behavior.
2. *Cooperation*: Both students are working, but they are working separately and independently, usually on different parts of the poster.
3. *Other*: At least one student is not working or contributes minimally to the task.

When the collaboration detector is used by a teacher who only cares that all students are participating, then the first two categories can be lumped together as "good" and the teacher only receives alerts about groups classified into the third category. On the other hand, if the teacher wants student to interact more intensely, then only the first category can be treated as "good" and the teacher receives alerts about groups classified into the second and third category.

This design matches the Chi's ICAP framework [1], which uses "Interactive" for category 1 (the I of ICAP). Category 2 means both students are "Constructive" (the C of ICAP). Category 3 means at least one student is Active, Passive or disengaged. The framework predicts that I > C > A > P for learning gains, which implies 1 > 2 > 3 for our categories. We used Chi's term "interactive" for the first category above.

Thus, we asked human judges to make the 3-way distinction above. However, during our analysis of accuracy, we considered the 3-way classification accuracy as well as several binary classifications obtained by lumping together 2 of the 3 categories.

Because our study was conducted in a camera-infused laboratory, we saw very little disengagement. Thus, the typical behavior of groups classified as "Other" was for one student to be doing all the work, while the other student watched and occasionally uttered brief agreements. In earlier work, we termed these categories "collaboration", "cooperation" and "asymmetric contribution" [21].

4 Analysis Methods

4.1 Audio Processing and Extraction of Audio Logs

The inputs to the collaboration detectors came only from the headset microphones; the other audio and video streams were seen only by the human coders. The first step in processing the headset mic audio was removal of background noise. Signal processing was carried out as proposed by Rafi et al. [18] along with few modifications. FFT windows was reduced to 0.25 from 0.5 and a soft mask by Wiener filtering method instead of a hard filter.

The binary version of the audio signal was obtained from the cleaned-up mic audio using a standard voice activity detector (WebRTC). The non-speech segments in the high quality audio signal was removed based on output from the voice activity detector.

4.2 Segmentation

As students answered the 12 questions, we noticed that they sometimes shifted their interaction patterns when transitioning from one question to the next. In order to avoid mixing up two different patterns of interactions, we placed segment boundaries between subtasks. More specifically, we used these criteria for placing a segment boundary: First, when there was a switch from one subtask to the other by any participant, a segment boundary was placed Second, if the particular activity took more than 4 min, then a segment boundary was placed immediately after the writing activity stopped for a brief time (like 5 s). This prevented overly long segments. Third, if a student went back to a different card and started writing, a segment boundary was placed. The average length of the segment was 110 s with standard deviation of 83 s.

4.3 Human Coding

Once the segmentation was performed, human annotators classified each segment as either interactive (I), cooperative (C) or other (O). The annotators used the audio-video stream obtained by the screen recorder and also used log data to understand various write events. If characteristics of multiple codes are found in the same segment, then the category with the greatest amount of time was assigned to the segment. Two human annotators tagged a sample of 80% of the overall segments. Inter rater agreement was considered acceptable with Cohen's kappa K = 0.76. For consistency across the whole dataset, the classifications of one annotator (the first author) were used in subsequent analysis.

4.4 Feature Extraction

In order to use standard machine learning algorithms to induce collaboration detectors for both the audio signal and the binary signal, the signals were represented as features.

The audio signal's features were from the OpenSMILE [20] audio feature set, which represents the state of the art in affect and paralinguistic recognition. Features were generated by a toolkit from the speech signal of each subject. The two individual subject feature vectors were then concatenated into one single feature vector for every segment.

The binary speech signal was characterized as a time series signal. This is obtained one per person. At a segment level, tsfresh [19] computes 794 time series features based on variations of both these signals with respect to time. The entire set of features obtained from tsfresh can be found in [19] and for OpenSMILE can be found in [20].

Group level features such as the duration of time when students spoke with each other (speech time per segment) and the duration of time when they did not speak (silence time) with each other were also extracted from the signal.

4.5 Feature Selection

Feature selection was performed because the number of features was greater than the number of observations. Pairwise correlations were performed on features likely to be redundant. Sets of highly correlated features (coefficient > 0.9) were reduced to a single feature chosen arbitrarily from the set.

For the high quality audio signal, recursive feature elimination is used to eliminate features that have low discriminative power across different classes.

For the binary signals, both the students' time series characterization along with its the collaboration class (I, C or O) was fed into tsFresh [19]. For each feature, it used Chi-square and other statistical tests to determine whether the features' value was reliability associated with the collaboration class. Only features whose p-value exceeded 0.05 were kept.

5 Results

As mentioned earlier, we developed collaboration detectors for several use cases. One pair of detectors distinguished all three categories (Interaction, Cooperation and Other). In addition, we generated binary classifiers focused on every single category.

5.1 Binary Classifier Focused on Cooperation

This section reports the accuracy of the binary classifier that was trained to discriminate Cooperation (C) from Non Cooperation (NC). We built classifiers using both high quality audio signal and the binary signal. Random Forests yielded the best results when compared to other algorithms such as logistic regression, bagging and boosting. The models were validated using tenfold cross validation. As Table 1 shows, the accuracy of the binary signal classifier (K = 0.53) was similar to the accuracy of the high quality audio classifier (K = 0.66).

Table 1. Confusion matrices for binary classifier focused on cooperation

High Quality Audio (F_1= 0.83, K= 0.66)

		Predictions	
		NC	C
True Class	NC	297	16
	C	15	38

Binary Logs (F_1= 0.76, K= 0.53)

		Predictions	
		NC	C
True Class	NC	294	19
	C	22	31

5.2 Binary Classifier Focused on Interaction

This section reports the accuracy of the binary classifier that trained to discriminate Interaction (I) from Non Interaction (NI). We built classifiers using both the high quality audio signal and the binary signal. Random Forests yielded the best results when compared to other algorithms. These models were validated using tenfold cross validation. As Table 2 shows, the accuracy of the binary signal classifier (K = 0.54) was similar to the accuracy of the high quality audio classifier (K = 0.62).

Table 2. Confusion Matrices of Binary Classifiers focused on Interaction

High Quality Audio (F_1= 0.80, K= 0.62)

		Predictions	
		I	NI
True Class	I	136	35
	NI	36	159

Binary Logs (F_1= 0.77, K= 0.54)

		Predictions	
		I	NI
True Class	I	129	42
	NI	40	155

5.3 Binary Classifier Focused on Other Category

This section reports the accuracy of the binary classifier that trained to discriminate Other (O) from Non Other (NO). We built classifiers using both high quality audio signal and the binary signal. Random Forests yielded the best results when compared to other algorithms. The models were validated using tenfold cross validation. As Table 3 shows, the accuracy of the binary signal classifier (K = 0.28) was similar to the accuracy of the high quality audio classifier (K = 0.36), but neither accuracy was high.

Table 3. Confusion Matrices for Binary Classifier Focused on Cooperation

High Quality Audio (F_1= 0.69, K= 0.36)

	Predictions		
		O	NO
True Class	O	82	60
	NO	50	174

Binary Logs (F_1= 0.64, K= 0.28)

		Predictions	
		O	NO
True Class	O	95	47
	NO	84	140

5.4 Ternary Classifier

This section reports the accuracy of the results of ternary classifier that is trained to discriminate three categories: Interaction (I), Cooperation (C) and Other (O). We built classifiers using both high quality audio signal and binary signal. Random Forests yielded the best results when compared to other types of algorithms. The models were validated using tenfold cross Validation. As Table 4 shows, the accuracy of the binary signal classifier (K = 0.44) was similar to the accuracy of the high quality audio signal classifier (K = 0.55).

Table 4. Confusion Matrices of Three way Classifiers

High Quality Audio (F_1 = 0.70, k=0.55)

		Predictions		
		I	O	C
True Class	I	94	31	17
	O	35	135	1
	C	17	0	36

Binary logs (F_1 = 0.63, k=0.44)

	Predictions			
		I	O	C
True Class	I	84	38	20
	O	46	123	2
	C	20	1	32

6 Discussion and Conclusion

When this project began, we did not think that detectors based on binary signals would perform well compared to collaboration detectors based on high quality audio signal. Against low expectations we got modest kappa scores - 0.53, 0.54, 0.44 except for the Other-focused binary classifier (0.28). To explain the Other classifier's inaccuracy, we can examine the results from the ternary classifier.

The contingency table of the ternary classifier (Table 4) shows that some of samples from the Interaction category are mistaken as Other category and vice versa. On the other hand, the Cooperation category was clearly separated from the Other category. This phenomenon is due to the fact that in some cases, when students worked alone with their partner watching, they continued to verbalize their writing and hence the machine learner assumed that they were Interacting with each other.

The binary classifier focused on Other category was minimally reliable since the classifier has to differentiate Other from Interaction and Cooperation combined together into one. The reliability was compromised since Interaction and Cooperation have entirely different audio characteristics and the Other category shares some characteristics of both. As a result, neither the binary nor full audio features could differentiate between them reliably. If the use case requires the Other category to be distinguished from non-Others, the ternary classifier should be used.

Although the collaboration detectors based on the binary signal performed reasonably well, there are a few caveats to consider. The first limitation is that synchronization of different data streams needs to be improved so that significant of manual labor can be avoided. Until this is achieved, our method cannot be used in classroom in real-time.

The second limitation is the usage of cards to automatically mark segment boundaries. This also allowed us to detect a change in subtask when subjects worked with each other. Although this helped in segmentation and improving reliability, we are not sure about the performance of the classifier when the boundaries are less salient.

Third, this study did not encounter off task behavior since it was a laboratory study recorded under a camera. Analysis of the semantic content of the speech may be necessary to detect off task behavior.

The fourth limitation is that the study was performed in artificial laboratory setting and it only involved two people at a time. This may have increased the accuracy of the full-audio classifier because there was no interference from other audio sources. This would also explain why Bassiou et al. [11] found that their collaboration detector based on a binary signal was more accurate than their collaboration detector based on a full audio signal: their speech was collected in a noisy classroom. Thus, their results combined with ours suggest a bright future for collaboration detection based on binary signals.

These finding suggest a clear direction for our future work, because they appear to solve several practical problems.

First, when a collaboration detection system needs high quality audio, the analog signal from the microphone must be sampled at a high bitrate. Transmitting these bitstreams wirelessly from 30 students can overload classroom radios, but using wires instead invites physical damage to the equipment when students or staff become entangled. In contrast, binary signals require less bandwidth, so 30 of them can probably be transmitted wirelessly reliably.

Finally, throat microphones can probably be used instead of headset microphones. These microphones detect noise only from the speaker and not ambient noise. Thus, the noise removal we performed prior to voice activity detection may not be necessary. Affordable throat microphones often produce somewhat distorted audio, so they would probably not work well for collaboration detection based on acoustic features.

Acknowledgements. This research was funded by the Diane and Gary Tooker chair for effective education in Science Technology Engineering and Math, by NSF grant IIS-1628782, and by the Bill and Melinda Gates Foundation under Grant OP1061281.

References

1. Chi, M.T.H., Wylie, R.: ICAP: a hypothesis of differentiated learning effectiveness for four modes of engagement activities. Educ. Psychol. **49**(4), 219–243 (2014)
2. Ladd, G.W., Kochenderfer-Ladd, B., Visconti, K.J., Ettekal, I., Sechler, C.M., Cortes, K.I.: Grade-school children’s social collaborative skills: links with partner preference and achievement. Am. Educ. Res. J. **51**(1), 152–183 (2014)
3. Magnisalis, I., Demetriadis, S., Karakostas, A.: Adaptive and intelligent systems for collaborative learning support: a review of the field. IEEE Trans. Learn. Technol. **4**(1), 5–20 (2011)
4. Martinez-Maldonado, R., Kay, J., Yacef, K.: An automatic approach for mining patterns of collaboration around an interactive tabletop. In: Lane, H.C., Yacef, K., Mostow, J., Pavlik, P. (eds.) AIED 2013. LNCS (LNAI), vol. 7926, pp. 101–110. Springer, Heidelberg (2013). https://doi.org/10.1007/978-3-642-39112-5_11
5. Bachour, K., Kaplan, F., Dillenbourg, P.: An interactive table for supporting participation balance in face-to-face collaborative learning. IEEE Trans. Learn. Technol. **3**(3), 203–213 (2010)
6. Gweon, G., Jain, M., McDonough, J., Raj, B., Rosé, C.P.: Measuring prevalence of other-oriented transactive contributions using an automated measure of speech style accommodation. Int. J. Comput. Support. Collaborative Learn. **8**(2), 245 (2013)
7. Gweon, G.: Assessment and support of the idea co-construction process that influences collaboration (2012)
8. Martinez-Maldonado, R., Dimitriadis, Y., Martínez-Monés, A., Kay, J., Yacef, K.: Capturing and analyzing verbal and physical collaborative learning interactions at an enriched interactive tabletop. Int. J. Comput. Support. Collaborative Learn. **8**(4), 455 (2013)
9. Martinez-Maldonado, R., Yacef, K., Kay, J.: TSCL: a conceptual model to inform understanding of collaborative learning processes at interactive tabletops. Int. J. Hum. Comput. Stud. **83**, 62–82 (2015)
10. Roman, F., Mastrogiacomo, S., Mlotkowski, D., Kaplan, F., Dillenbourg, P.: Can a table regulate participation in top level managers’ meetings? In: Proceedings of the 17th ACM International Conference on Supporting Group Work - GROUP 2012 (2012)
11. Bassiou, N., et al.: Privacy-preserving speech analytics for automatic assessment of student collaboration. In: Proceedings Interspeech, pp. 888–892 (2016)

12. Agrawal, P., Udani, M.: The automatic assessment of knowledge integration processes in project teams, Long papers, p. 462 (2011)
13. Gweon, G., Jain, M., McDonogh, J., Raj, B.: Predicting idea co-construction in speech data using insights from sociolinguistics. In: Proceedings of the Learning Sciences: The Future of Learning (2012)
14. http://map.mathshell.org/index.php
15. Meier, A., Spada, H., Rummel, N.: A rating scheme for assessing the quality of computer-supported collaboration processes. Comput.-Supported Collaborative Learn. **2**, 63–86 (2007)
16. Kahrimanis, G., Chounta, I.-A., Avouris, N.: Validating empirically a rating approach for quantifying the quality of collaboration. In: Daradoumis, T., Demetriadis, S., Xhafa, F. (eds.) Intelligent Adaptation and Personalization Techniques, pp. 295–310. Springer, Heidelberg (2012). https://doi.org/10.1007/978-3-642-28586-8_13
17. Gweon, G., Jain, M., McDonough, J., Raj, B., Rose, C.P.: Measuring prevalence of other-oriented transactive contributions using an automated measure of speech style accommodation. Int. J. Comput. Support. Collaborative Learn. **8**(2), 245–265 (2013)
18. Rafii, Z., Pardo, B.: Music/Voice separation using the similarity matrix, pp. 583–588 (2012)
19. Christ, M., Braun, N., Neuffer, J., Kempa-Liehr, A.W.: Time series FeatuRe extraction on basis of scalable hypothesis tests (tsfresh–A Python package). Neurocomputing (2018)
20. Eyben, F., Wöllmer, M., Schuller, B.: OpenSMILE: the Munich versatile and fast open-source audio feature extractor, pp. 1459–1462. ACM (2010)
21. Viswanathan, S.A., VanLehn, K.: Using the tablet gestures and speech of pairs of students to classify their collaboration. IEEE Trans. Learn. Technol. **11**, 230–242 (2018)
22. Cheema, S., VanLehn, K., Burkhardt, H., Pead, D., Schoenfeld, A.: Electronic posters to support formative assessment. In: Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems, pp. 1159–1164 (2016)