The Challenge of Noisy Classrooms: Speaker Detection During Elementary Students' Collaborative Dialogue

Yingbo Ma[1] (✉), Joseph B. Wiggins[1], Mehmet Celepkolu[1], Kristy Elizabeth Boyer[1] (✉), Collin Lynch[2], and Eric Wiebe[2]

[1] University of Florida, Gainesville, FL 32601, USA
{yingbo.ma, jbwiggi3, mckolu, keboyer}@ufl.edu
[2] North Carolina State University, Raleigh, NC 27606, USA
{cflynch, wiebe}@ncsu.edu

Abstract. Adaptive and intelligent collaborative learning support systems are effective for supporting learning and building strong collaborative skills. This potential has not yet been realized within noisy classroom environments, where automated speech recognition (ASR) is very difficult. A key challenge is to differentiate each learner's speech from the background noise, which includes the teachers' speech as well as other groups' speech. In this paper, we explore a multimodal method to identify speakers by using visual and acoustic features from ten video recordings of children pairs collaborating in an elementary school classroom. The results indicate that the visual modality was better for identifying the speaker when in-group speech was detected, while the acoustic modality was better for differentiating in-group speech from background speech. Our analysis also revealed that recurrent neural network (RNN)-based models outperformed convolutional neural network (CNN)-based models with higher speaker detection F-1 scores. This work represents a critical step toward the classroom deployment of intelligent systems that support collaborative learning.

Keywords: Adaptive and Intelligent Collaborative Learning Support · Classroom Environment · Speaker Detection · Multimodal Learning

1 Introduction

Adaptive and intelligent collaborative learning support (AICLS) systems [25] provide personalized feedback [45] to individual students working in pairs or groups. An AICLS system not only analyzes the group interaction [27] and provides tailored supports during the problem-solving process [36], but also adapts its content presentation or navigation support according to the learners' collaboration activity. This technology has been shown to be effective for improving students' learning outcomes [1], increasing their engagement in learning [46] and helping students build strong collaboration skills [26]. Early results have shown that adaptive supports are better than non-adaptive supports for providing flexible guidance [17] and improving learning outcomes [21].

Despite this promise, AICLS cannot currently support real-time collaboration between children speaking together in noisy classrooms. Instead, most current AICLS systems are designed for remote/distributed collaboration settings where an individual's speech and actions can easily be isolated [31,41] and where students' individual learning activities are often identified through analyzing students' log actions [37] or the group discourse through textual chat [41,44,45]. This is in part because classrooms are noisy: they feature multiple overlapping audio sources, and deploying ASR in these environments is difficult due to the challenges of handling background noise and detecting/isolating speech and speakers [3]. This problem could be mitigated with students wearing their own microphone with noise cancelling capabilities; however, most schools are unable to afford deploying these devices en masse. In addition, headsets detract from the fluid interplay of individual, small group, and whole class discourse.

To address these challenges and move toward AICLS systems that are viable for use in noisy classrooms, one task that must be addressed is detecting which child from a collaborating pair is speaking at a given moment. This paper reports on a novel speaker detection method that uses visual and acoustic features from video recordings of learners collaborating, with the goal of identifying which child is speaking. The proposed approach analyzes a single mixed audio source from two students in the group, which does not require their audios to be recorded into separate channels. In addition, the approach utilizes visual features detected from two children's faces, which could act as supplementary indicators to acoustic features. To the best of our knowledge, this paper presents the first empirical evaluation combining visual and acoustic features on the challenging task of identifying the speaker within student pairs in noisy classroom contexts.

2 Related Work

Recently, AICLS systems have been deployed for various learning domains, such as computer science learning [47], medical training [9], and music learning [24]. In this section we focus on AICLS systems within the context of computer science education. Current systems have used a variety of methods to identify each student's activity during collaboration. SIENA [29] tracked individual's learning progress by calculating the learner's posterior knowledge after he/she answered a question. NUCLEO [37] built an adaptation model for each learner based on the individual score obtained among team partners and the system-user interaction process, such as number of files created and answered messages. SCEPPSys [41] and Peer Tutor [45] analyzed group discourse from students' textual chat history. CycleTalk Chat [21] identified individuals by assigning each student an audio-based chat client and collecting their dialogues separately.

The aforementioned systems were all designed for remote/distributed collaboration where students were not co-located. There have been very few systems that analyze student dialogues while they are working collaboratively in person.

Harsley et al. [12] designed Collab-ChiQat for analyzing student activity during collaboration; yet, the system required students to self-report who authored each line of code. Yett et al. [47] analyzed individual log actions of co-located students participating in a collaborative programming environment. The authors suggested that future work should combine log-based analysis and discourse analysis, which relies on the accurate differentiation of speech between individuals within the group. In another classroom study, Celepkolu et al. [5] designed a visualization tool to help individual students reflect on their collaborative dialogues. Even though the tool automatically analyzed the dialogue and generated the visualizations based on the transcriptions, it still required the dialogue to be manually transcribed. Blanchard et al. [3] tested and compared five ASR engines such as Google Speech and Bing Speech with audio data collected in middle school classrooms, but their focus was on teachers who wore individual wireless microphones. Li et al. [22] designed a Siamese neural network to detect dialogues for teachers and students in both *online* and *offline* classroom audio recordings. Although a promising level of speaker detection was achieved, the authors suggested that future work should combine both audio and video data.

Our study differs from these studies in three ways: First, our dataset consisted of pairs of students sitting next to each other, sharing the computer, with background noise from other students and the teacher. Such a research context makes distinguishing the speakers much more challenging. Second, we identify speakers by using video recordings (audio and video images) of the students' collaborative interaction process collected from the built-in computer webcam (without any headsets). Third, we applied recent machine learning techniques and compared the performance between CNN-based models and RNN-based models. Prior work by Hu et al. [15] has shown promise using CNN-based models to localize and identify each speaking character in a TV/movie/live show video, but did not consider the natural temporal connections within the sequential data. In contrast, RNN-based models represent a novel approach to solving this problem and have been used by Soleymani et al. [39] to analyze a speaker's verbal and nonverbal behaviors associated with self-disclosure with multimodal features extracted from video, audio and text data.

3 Data

3.1 Collection and Preprocessing

Our dataset was collected from 20 children (10 pairs) in 4th/5th grade classrooms in an elementary school in the southeastern United States in 2019. Among the children, 9 identified themselves as females and 11 as males. The students collaborated on a series of coding activities, in which they learned fundamental CS concepts such as variables, conditionals, and loops using Netsblox [30], a block-based learning environment. Each group's collaboration process was videotaped by the front-facing camera of their computer; meanwhile, the audio was recorded by the computer microphone without any additive noise cancellation equipment. The corpus contained a total of 7 hours and 22 minutes of video recordings; raw audio recordings were then extracted from video recordings using FFmpeg [10], an open-source video converter.

Since noise sensitivity is a significant challenge for speech-related tasks, we approximated the quality of our audio recordings by following the method used by Tan et al. [40] to compute the *posterior* signal-to-noise ratio (SNR), the logarithmic ratio of the energy of the noisy speech to the energy of the noise. The average estimated SNR over ten recordings was +2.20 dB (as shown in Table 1), indicating a *fair* audio quality. Howard et al. [14] reported that the typical classroom SNRs range from −7 dB to +5 dB, while an SNR of +15 dB or above indicates *good* speech quality.

3.2 Annotation

We used ELAN [7] to synchronize video and audio clips and annotate them. The data was tagged at a one-second granularity, a time window previously used in similar acoustic classification tasks [42]. We tagged each one-second clip in one of three ways: Left Child (the child sitting in the left of the video was speaking), Right Child (the child sitting on the right was speaking), and Silence (neither Left Child nor Right Child was speaking). No children switched position during the activity. When the clip contained overlapping speech, we tagged the clip based on which child's speech was more audible. Table 1 shows the details of the corpus. The first author annotated the first four of ten videos, and the remaining six videos were annotated by three other annotators. To measure the labeling reliability, the first author then independently tagged 10% continuous video excerpts from the data tagged by other annotators. The Cohen's kappa scores between the first author and each of our three annotators were 0.8521, 0.7109, 0.7526 respectively, indicating substantial inner-annotator agreement [4].

Table 1: Details of the collected classroom recording corpus

Video ID	**Duration(second)**	**Left Child(second)**	**Right Child(second)**	**Silence(second)**	**SNR(dB)**
1	2574	569	386	1619	+2.60
2	2199	394	498	1307	+1.24
3	2550	397	605	1548	+1.57
4	2693	416	485	1792	+3.17
5	3019	617	314	2088	+3.35
6	2665	165	302	2198	+2.67
7	2940	311	426	2203	+2.48
8	2804	526	275	2003	+2.12
9	2350	344	509	1497	+1.40
10	2673	377	604	1692	+1.39
In total	26467	4116 (15.55%)	4404 (16.64%)	17947 (67.81%)	+2.20

[1] Silence class also includes the clips in which the in-group children were silent but background speech was detected from the teacher and other children in the classroom. Background speech was irrelevant to in-group interaction and should not be taken into consideration for further in-group interaction analysis by the AICLS system

4 Features

4.1 Visual Feature: Dense Optical Flow

Facial movements, especially around the lip area, are critical to detect speakers [8,35]. In this paper, we extracted visual features using the dense optical flow [38] from children's faces in each pair (see Fig 1). Dense optical flow uses the variation of pixels to calculate the object motion gradient along time. To compute the dense optical flows for two children in the video, we first extracted their faces using the deep learning-based face detector [33] from OpenCV [32] (a real-time computer vision library). Then, we re-scaled all faces into the same image size and used the *cv2.calcOpticalFlowFarneback()* [34] function from OpenCV to calculated one dense optical flow on their faces for each second. We applied dense optical flow on the whole face instead of the mouth region because whole-face optical flows were more robust to instances in which the child was not directly facing the camera, or in the case of low-resolution recording. Dense optical flow images were generated in grey-scale because the color in dense optical flow denotes the movement direction, which was not needed to identify speakers.

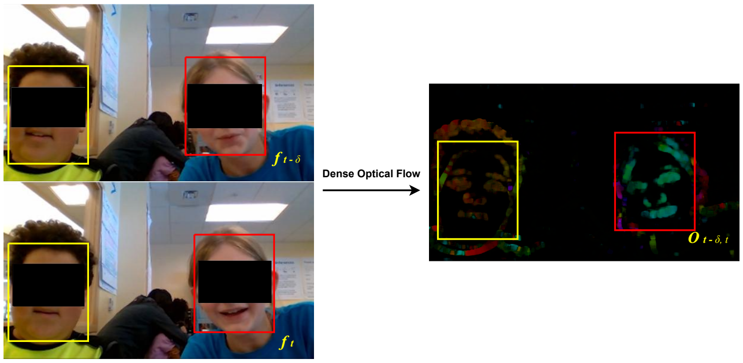

Fig. 1: *Left*: two sample frames ($f_{t-\delta}$ and f_t) from a one-second video clip when the left child was not speaking and the right child was speaking. *Right*: dense optical flow $O_{t-\delta,t}$, which represents the motion detected between the two frames. In this case, more motion was detected from the right child's face: the intensity in the dense optical flow denotes movement speed.

4.2 Acoustic Feature: Mel Spectrogram

We converted each one-second audio clip into one mel spectrogram, an image representation that describes an audio's time-frequency distribution where the frequencies are converted in the mel scale—a perceptual scale of pitches judged by human listeners. One advantage of the mel spectrogram over traditional acoustic features [19] (pitch, energy, mfcc coefficients, etc.) is it shows the variance of acoustic frequency and energy over time, which is useful for analyzing sequential data. In an mel spectrogram, the x-axis represents time and the y-axis represents frequency. We generated mel spectrograms using *librosa* [28], a python library for audio analysis. Figure 2 shows four mel spectrograms generated from 4 different audio clips.

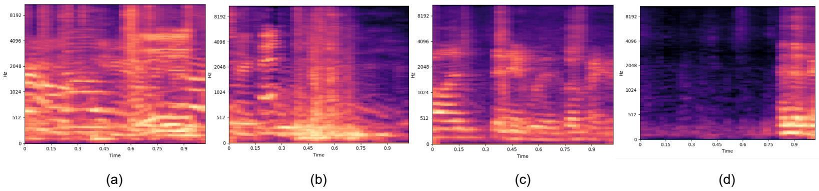

Fig. 2: Four mel spectrograms: *(a)*: The target child spoke over the whole audio clip; *(b)*: The speech from the target child only presented in the first 0.6 second of the audio clip; *(c)*: Silence, the target children were silent but background speech from the teacher or other children was audible; *(d)*: Silence, none speech detected over the whole audio clip

5 Methods: Speaker Detection

In this study, we conducted two experiments. First, to analyze the performance of different feature combinations, we compared the results of uni-modality (only visual or only acoustic features) with multi-modality (visual and acoustic features). Second, to compare the performance of different model architectures, we tested our dataset with two types of commonly used models (CNN-based and RNN-based).

Experiment 1: Comparing Uni-modality with Multi-modality. Figure 3 shows the high-level structure of the multimodal learning model, which was divided into three parallel streams (one visual stream for the left child, one visual stream for the right child, and one acoustic stream for both children). The model consisted of two parts: a modality encoding network and a sequence-based recurrent network. Since the visual and acoustic feature representations are both images, we used CNN-based models in the modality encoding network. We used ResNet-50 [13], a pre-trained CNN-based model that achieved the highest image classification accuracy on ImageNet [16], to map each image representation into a feature embedding. In the second sequence-based recurrent network, we used Bi-directional Long Short-Term Memory (Bi-LSTM) to learn temporal dependencies between sequential feature representations. We tested different time steps of the Bi-LSTM from 2 to 5. Each output of the Bi-LSTM is a feature embedding followed by a *softmax* layer [23] to calculate the class scores. Finally, we combined the class scores from three separate streams by averaging fusion [38]. The model (Code available on GitHub [1]) was implemented in Python with the Keras [18] API. Two visual streams were used for evaluating the performance of the visual modality, and one acoustic stream was used for evaluating the performance of the acoustic modality.

Experiment 2: Comparing CNN-based models with RNN-based models. We selected two types of commonly used models (CNN-based and RNN-based) that were proposed in the recent literature. Hu et al. [15] proposed a two-stream CNN-based learning framework for localizing and identifying each speaking character in a TV/movie/live show video. The model used convolutional layers as face feature extractors, then learned a unified multimodal classi-

[1] https://github.com/yingbo-ma/The-Challenge-of-Noisy-Classrooms-AIED2021

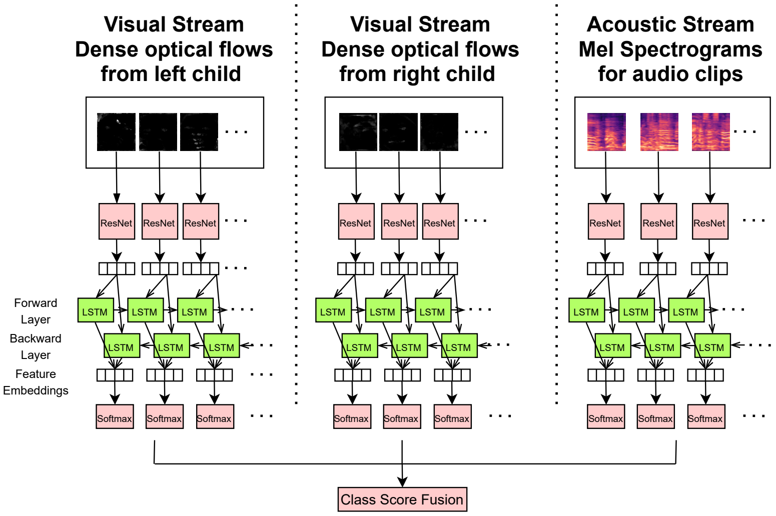

Fig. 3: Multimodal learning model

fier with fusion features combined from visual and acoustic features. Soleymani et al. [39] proposed a ResNet + GRU (gated recurrent unit)-based method to analyze a speaker's verbal and nonverbal behaviors associated with self-disclosure. The model built three separate classifiers based on the visual features extracted with ResNet, the acoustic features extracted with VGGish [43] (a pre-trained acoustic feature extractor trained on audio spectrograms), and the language features extracted with BERT [6] (a pre-trained language model that can map the spoken utterances to feature representations). The model then performed late fusion by simply averaging the output from all modalities. Since the feature fusion strategy was not the focus of this work, we implemented the above-mentioned models followed by the description of the model architecture in the two papers, and still used late averaging fusion. The CNN model [15] consisted of three stacked convolution + pooling layers followed by a fully connected layer. The RNN model [39] consisted of the pre-trained ResNet-50 [13] followed by a single GRU layer with 128 hidden units.

During the model training process across the two experiments, we conducted experiments on each video recording and used ten-fold cross-validation to train and evaluate the model. The network updated weights with an Adam optimizer [20] with the learning rate of 0.0001. We evaluated the trained model with the F-1 score [11] combined from precison and recall for each one of the three classes. Although F-1 score can be used as a general measurement of model performance, including precision and recall provides additional information. The context of collaborative dialogue may shift the cost of false negatives versus false positives, so these additional scores allows us to weigh each case.

6 Results

Results for Experiment 1 Figure 4 shows the performance of uni-modality and multi-modality. In Figure 4-Left, the acoustic modality outperformed the visual modality and the combined modality when identifying the Silent class. In Figure 4-Middle and Figure 4-Right, the visual modality outperformed the acoustic modality and the combined modality when identifying one of the speech classes. Table 2 compares the different modalities with averaged precision, recall and F-1 score for each class. Table 3 displays results for the multimodal learning model's performance with different time steps. The time step of 3 performed the best for classifying the Silence class and the time step of 3 and 4 both performed similarly well at classifying the Left and Right Child class.

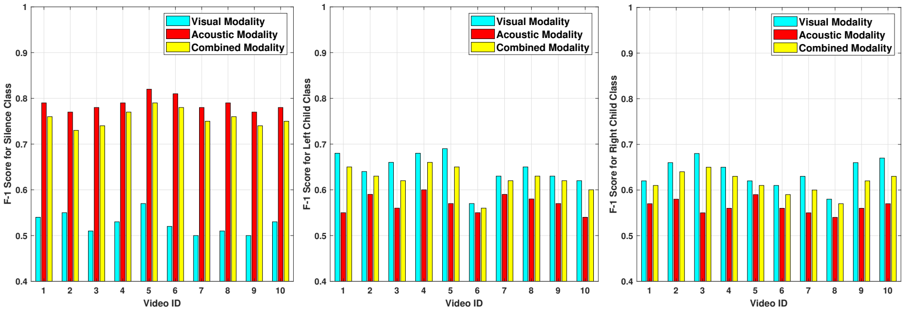

Fig. 4: F-1 score for three classes across corpus using different modalities. *Left*: Silence class—acoustic modality outperformed visual modality; *Middle*: Left Child class—visual modality outperformed acoustic modality; *Right*: Right Child class—visual modality outperformed acoustic modality

Table 2: Performance for each class under different modalities

	Silence			Left Child			Right Child		
	Precision	Recall	F-1 Score	Precision	Recall	F-1 Score	Precision	Recall	F-1 Score
Visual Modality	0.55	0.49	0.52	0.59	0.69	0.64	0.59	0.68	0.63
Acoustic Modality	0.68	0.89	0.78	0.68	0.50	0.56	0.68	0.49	0.55
Combined Modality	0.73	0.79	0.76	0.66	0.61	0.63	0.66	0.60	0.62

Table 3: Performance of the multimodal learning model with different time steps in Bi-LSTM

Time Step	2			3			4			5		
	Precision	Recall	F-1 Score	Precision	Recall	F-1 Score	Precision	Recall	F-1 Score	Precision	Recall	F-1 Score
Silence	0.72	0.84	0.76	0.72	0.84	0.77	0.72	0.84	0.77	0.71	0.83	0.76
Left Child	0.63	0.58	0.60	0.66	0.60	0.63	0.66	0.59	0.62	0.64	0.58	0.61
Right Child	0.64	0.59	0.61	0.66	0.60	0.62	0.67	0.60	0.63	0.67	0.60	0.63

Results for Experiment 2 Table 4 shows the performance of different models on our corpus. The CNN architecture [15] performed the best at classifying Silence; Both the ResNet + GRU model [39] and the ResNet + Bi-LSTM model in our paper performed similarly, with better classification performance on Left Child and Right Child than the CNN architecture. ResNet + Uni-LSTM performed comparably with ResNet + Bi-LSTM, potentially indicating that whether a child intends to speak has stronger connection with his/her prior dialogues than latter dialogues.

Table 4: Performance of different models

Model	Class	Precision	Recall	F-1 Score
CNN [15]	Silence	0.70	0.87	0.78
	Left Child	0.65	0.53	0.58
	Right Child	0.65	0.53	0.58
ResNet + GRU [39]	Silence	0.72	0.81	0.76
	Left Child	0.67	0.58	0.62
	Right Child	0.66	0.58	0.62
ResNet + Uni-directional LSTM	Silence	0.72	0.83	0.77
	Left Child	0.66	0.60	0.63
	Right Child	0.65	0.60	0.62
ResNet + Bi-directional LSTM	Silence	0.72	0.84	0.77
	Left Child	0.66	0.60	0.63
	Right Child	0.66	0.61	0.62

7 Discussion

This work evaluated several unimodal and multimodal learning frameworks' performance on identifying the speaker within pairs of children in a noisy elementary school classroom. Our results show the effectiveness of using visual optical flow and acoustic mel spectrogram for this task, and achieved averaged F-1 scores of 0.76 for Silence, 0.63 for Left Child, and 0.62 for Right Child.

These results have several implications for developing AICLS systems that can be utilized for personalized supports during collaborative learning in noisy classrooms. In the experiment investigating the contribution of each modality, the results showed that only using the visual modality yielded a higher F-1 score on detecting speakers compared to using the combined visual and acoustic modality. However, only using the visual modality has potential drawbacks due to lower Precision and higher Recall, meaning the model falsely reported more irrelevant background speech samples as in-group speech samples. This could potentially be misleading for an AICLS system. For example, the system might provide support when students are listening to teacher's lecture because the system would falsely classify the teacher's dialogues as the students' dialogues. Therefore, the feature modality should be carefully selected based on the noise level of a classroom. If the classroom is relatively quiet, using the visual modality may provide better speaker detection performance and report more true in-group speech samples. However, if a classroom is noisy and the in-group speech is overwhelmed by the background speech, the results suggest using the combined visual and acoustic modality may help. The experiment of comparing CNN-based models and RNN-based models showed that the CNN-based model performed better in differentiating in-group speech from background noise, and RNN-based models performed better for distinguishing between in-group speakers. Compared to silence, speech tends to have a more temporal connection, which was better modeled by the sequential neural network of the RNN. Therefore, CNN-based models would be better to use when the proportion of speech is much lower than the proportion of silence in students' dialogues, and RNN-based models would be more appropriate to use when in-group children interact with partners more frequently.

There are important limitations of the present approach. First, although our framework achieved promising results, the generalizability of the model has not been shown. Each learned model depends on the unique audio characteristics of children in the training set. In addition, the effectiveness of visual features largely relies on the correct setup of the video data collection process. The speaker detection performance on videos 6 and 8 was much lower than the averaged results across because in both videos, the front-facing camera was not positioned correctly. On the other hand, the effectiveness of acoustic feature depends on the group members' voices and the audio quality. If the frequency range of two children's voices is narrow (this often happens when two children in the group are of the same gender), the performance of using acoustic features would deteriorate.

8 Conclusions and Future Work

AI to support collaborative learning in classrooms holds great promise, but the tasks of identifying who is speaking, and what they are saying, present great challenges. This paper investigated the task of speaker detection. By utilizing features from the visual modality and the acoustic modality, our RNN-based model achieved encouraging speaker detection performance. The results indicated that the acoustic modality performed better at differentiating in-group speech and background noise; and the visual modality performed better in identifying in-group speakers. We also compared the performance of different models on this task and found that RNN-based models outperformed CNN-based models in modeling the temporal connection within the speech.

These results highlight several directions for future work. First, while the features used in this paper were promising, other features should be investigated (e.g., lip motion tracking, linguistic features). In addition, performance of cloud-based ASR services needs to be tested as well as the use of other popular face detection toolkits (e.g., Openface 2.0 [2]), and the results of learning models with different fusion strategies (feature versus class score fusion) needs further analysis. The work reported in this paper was a first step toward building an intelligent collaboration support system that can detect interactions between a pair of children and provide adaptive supports during learning within the noisy classroom environment. As we move toward this goal, we will be able to build and investigate systems that can significantly improve children's collaborative learning experience in classrooms.

9 Acknowledgement

This research is supported by the National Science Foundation through grant DRL-1721160. Any opinions, findings, conclusions, or recommendations expressed in this report are those of the participants, and do not necessarily represent the official views, opinions, or policy of the National Science Foundation.

References

1. Ahmed, I., Mawasi, A., Wang, S., Wylie, R., Bergner, Y., Whitehurst, A., Walker, E.: Investigating help-giving behavior in a cross-platform learning environment. In: Proceedings of the International Conference on Artificial Intelligence in Education. pp. 14–25. Springer (2019)
2. Baltrusaitis, T., Zadeh, A., Lim, Y.C., Morency, L.P.: Openface 2.0: Facial behavior analysis toolkit. In: Proceedings of the International Conference on Automatic Face & Gesture Recognition. pp. 59–66. IEEE (2018)
3. Blanchard, N., Brady, M., Olney, A.M., Glaus, M., Sun, X., Nystrand, M., Samei, B., Kelly, S., D'Mello, S.: A study of automatic speech recognition in noisy classroom environments for automated dialog analysis. In: Proceedings of the International Conference on Artificial Intelligence in Education. pp. 23–33. Springer (2015)
4. Brack, A., D'Souza, J., Hoppe, A., Auer, S., Ewerth, R.: Domain-independent extraction of scientific concepts from research articles. In: Proceedings of the European Conference on Information Retrieval. pp. 251–266. Springer (2020)
5. Celepkolu, M., Wiggins, J.B., Galdo, A.C., Boyer, K.E.: Designing a visualization tool for children to reflect on their collaborative dialogue. International Journal of Child-Computer Interaction **27**, 100232 (2021)
6. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)
7. ELAN: https://archive.mpi.nl/tla/elan
8. Ellamil, M., Susskind, J.M., Anderson, A.K.: Examinations of identity invariance in facial expression adaptation. Cognitive, Affective, & Behavioral Neuroscience **8**(3), 273–281 (2008)
9. Fadljević, L., Maitz, K., Kowald, D., Pammer-Schindler, V., Gasteiger-Klicpera, B.: Slow is good: the effect of diligence on student performance in the case of an adaptive learning system for health literacy. In: Proceedings of the Tenth International Conference on Learning Analytics & Knowledge. pp. 112–117 (2020)
10. FFmpeg: https://github.com/FFmpeg/FFmpeg
11. Goutte, C., Gaussier, E.: A probabilistic interpretation of precision, recall and f-score, with implication for evaluation. In: Proceedings of the European Conference on Information Retrieval. pp. 345–359. Springer (2005)
12. Harsley, R., Green, N., Di Eugenio, B., Aditya, S., Fossati, D., Al Zoubi, O.: Collabchiqat: A collaborative remaking of a computer science intelligent tutoring system. In: Proceedings of the 19th ACM Conference on Computer Supported Cooperative Work and Social Computing Companion. pp. 281–284 (2016)
13. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 770–778 (2016)
14. Howard, C.S., Munro, K.J., Plack, C.J.: Listening effort at signal-to-noise ratios that are typical of the school classroom. International Journal of Audiology **49**(12), 928–932 (2010)
15. Hu, Y., Ren, J.S., Dai, J., Yuan, C., Xu, L., Wang, W.: Deep multimodal speaker naming. In: Proceedings of the 23rd ACM International Conference on Multimedia. pp. 1107–1110 (2015)
16. ImageNet: http://www.image-net.org/

17. Karakostas, A., Demetriadis, S.: Enhancing collaborative learning through dynamic forms of support: the impact of an adaptive domain-specific support strategy. Journal of Computer Assisted Learning **27**(3), 243–258 (2011)
18. Keras: https://keras.io/api/
19. Kiktova, E., Lojka, M., Pleva, M., Juhar, J., Cizmar, A.: Comparison of different feature types for acoustic event detection system. In: Proceedings of the International Conference on Multimedia Communications, Services and Security. pp. 288–297. Springer (2013)
20. Kingma, D., Ba, J.: Adam: A method for stochastic optimization. Proceedings of the International Conference on Learning Representations (12 2014)
21. Kumar, R., Rosé, C.P., Wang, Y.C., Joshi, M., Robinson, A.: Tutorial dialogue as adaptive collaborative learning support. Frontiers in Artificial Intelligence and Applications **158**, 383 (2007)
22. Li, H., Wang, Z., Tang, J., Ding, W., Liu, Z.: Siamese neural networks for class activity detection. In: Proceedings of the International Conference on Artificial Intelligence in Education. pp. 162–167. Springer (2020)
23. Liu, W., Wen, Y., Yu, Z., Yang, M.: Large-margin softmax loss for convolutional neural networks. In: Proceedings of the International Conference on Machine Learning. pp. 507–516 (2016)
24. Lyu, F., Tian, F., Feng, W., Cao, X., Zhang, X., Dai, G., Wang, H.: Ensewing: creating an instrumental ensemble playing experience for children with limited music training. In: Proceedings of the CHI Conference on Human Factors in Computing Systems. pp. 4326–4330 (2017)
25. Magnisalis, I., Demetriadis, S., Karakostas, A.: Adaptive and intelligent systems for collaborative learning support: A review of the field. IEEE transactions on Learning Technologies **4**(1), 5–20 (2011)
26. Marcos-García, J.A., Martínez-Monés, A., Dimitriadis, Y.: Despro: A method based on roles to provide collaboration analysis support adapted to the participants in cscl situations. Computers & Education **82**, 335–353 (2015)
27. Martínez-Monés, A., Harrer, A., Dimitriadis, Y.: An interaction-aware design process for the integration of interaction analysis into mainstream cscl practices. In: Analyzing Interactions in CSCL, pp. 269–291. Springer (2011)
28. McFee, B., Raffel, C., Liang, D., Ellis, D.P., McVicar, M., Battenberg, E., Nieto, O.: librosa: Audio and music signal analysis in python. In: Proceedings of the 14th Python in Science Conference. vol. 8, pp. 18–25 (2015)
29. Moreno, L., Popescu, B., Groenwald, C.: Teaching computer architecture using a collaborative approach: the siena tool tutorial sessions and problem solving. Learning **2**, 10 (2013)
30. Netsblox: https://netsblox.org/
31. Nguyen, V., Dang, H.H., Do, N.K., Tran, D.T.: Enhancing team collaboration through integrating social interactions in a web-based development environment. Computer Applications in Engineering Education **24**(4), 529–545 (2016)
32. OpenCV: https://github.com/opencv/opencv
33. OpenCV-Face-Detector: https://github.com/opencv/opencv/tree/master/samples/dnn/face_detector
34. OpenCV-Optical-Flow: https://opencv-python-tutroals.readthedocs.io/en/latest/py_tutorials/py_video/py_lucas_kanade/py_lucas_kanade.html
35. Ren, J., Hu, Y., Tai, Y.W., Wang, C., Xu, L., Sun, W., Yan, Q.: Look, listen and learn—a multimodal lstm for speaker identification. In: Proceedings of the AAAI Conference on Artificial Intelligence. pp. 3581–3587 (2016)

36. Rodríguez, F.J., Boyer, K.E.: Discovering individual and collaborative problem-solving modes with hidden markov models. In: Proceedings of the International Conference on Artificial Intelligence in Education. pp. 408–418. Springer (2015)
37. Sancho, P., Fuentes-Fernández, R., Fernández-Manjón, B.: Nucleo: Adaptive computer supported collaborative learning in a role game based scenario. In: Proceedings of the IEEE International Conference on Advanced Learning Technologies. pp. 671–675. IEEE (2008)
38. Simonyan, K., Zisserman, A.: Two-stream convolutional networks for action recognition in videos. Advances in Neural Information Processing Systems **27**, 568–576 (2014)
39. Soleymani, M., Stefanov, K., Kang, S.H., Ondras, J., Gratch, J.: Multimodal analysis and estimation of intimate self-disclosure. In: Proceedings of the International Conference on Multimodal Interaction. pp. 59–68 (2019)
40. Tan, Z.H., Lindberg, B.: Low-complexity variable frame rate analysis for speech recognition and voice activity detection. IEEE Journal of Selected Topics in Signal Processing **4**(5), 798–807 (2010)
41. Tsompanoudi, D., Satratzemi, M., Xinogalos, S.: Evaluating the effects of scripted distributed pair programming on student performance and participation. IEEE Transactions on education **59**(1), 24–31 (2015)
42. Varatharaj, A., Botelho, A.F., Lu, X., Heffernan, N.T.: Supporting teacher assessment in chinese language learning using textual and tonal features. In: Proceedings of the International Conference on Artificial Intelligence in Education. pp. 562–573. Springer (2020)
43. VGGish: https://github.com/tensorflow/models/tree/master/research/audioset/vggish
44. Vizcaíno, A., Contreras, J., Favela, J., Prieto, M.: An adaptive, collaborative environment to develop good habits in programming. In: Proceedings of the International Conference on Intelligent Tutoring Systems. pp. 262–271. Springer (2000)
45. Walker, E., Rummel, N., Koedinger, K.R.: Adaptive intelligent support to improve peer tutoring in algebra. Proceedings of the International Conference on Artificial Intelligence in Education **24**(1), 33–61 (2014)
46. Walker, E., Rummel, N., Koedinger, K.R., et al.: Modeling helping behavior in an intelligent tutor for peer tutoring. In: Proceedings of the International Conference on Artificial Intelligence in Education. pp. 341–348 (2009)
47. Yett, B., Hutchins, N., Snyder, C., Zhang, N., Mishra, S., Biswas, G.: Evaluating student learning in a synchronous, collaborative programming environment through log-based analysis of projects. In: Proceedings of the International Conference on Artificial Intelligence in Education. pp. 352–357. Springer (2020)