Detecting Impasse During Collaborative Problem Solving with Multimodal Learning Analytics

Yingbo Ma
University of Florida
Gainesville, USA
yingbo.ma@ufl.edu

Mehmet Celepkolu
University of Florida
Gainesville, USA
mckolu@ufl.edu

Kristy Elizabeth Boyer
University of Florida
Gainesville, USA
keboyer@ufl.edu

ABSTRACT

Collaborative problem solving has numerous benefits for learners, such as improving higher-level reasoning and developing critical thinking. While learners engage in collaborative activities, they often experience *impasse*, a potentially brief encounter with differing opinions or insufficient ideas to progress. Impasses provide valuable opportunities for learners to critically discuss the problem and re-evaluate their existing knowledge. Yet, despite the increasing research efforts on developing multimodal modeling techniques to analyze collaborative problem solving, there is limited research on detecting impasse in collaboration. This paper investigates multimodal detection of impasse by analyzing 46 middle school learners' collaborative dialogue—including speech and facial behaviors—during a coding task. We found that the semantics and speaker information in the linguistic modality, the pitch variation in the audio modality, and the facial muscle movements in the video modality are the most significant unimodal indicators of impasse. We also trained several multimodal models and found that combining indicators from these three modalities provided the best impasse detection performance. To the best of our knowledge, this work is the first to explore multimodal modeling of impasse during the collaborative problem solving process. This line of research contributes to the development of real-time adaptive support for collaboration.

CCS CONCEPTS

• **Human-centered computing → Collaborative and social computing**.

KEYWORDS

Collaborative Problem Solving; Pair Programming; Impasse Detection; Multimodal Learning Analytics

ACM Reference Format:
Yingbo Ma, Mehmet Celepkolu, and Kristy Elizabeth Boyer. 2022. Detecting Impasse During Collaborative Problem Solving with Multimodal Learning Analytics. In *LAK22: 12th International Learning Analytics and Knowledge Conference (LAK22), March 21–25, 2022, Online, USA.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3506860.3506865

LAK22, March 21–25, 2022, Online, USA

ACM ISBN 978-1-4503-9573-1/22/03…$15.00
https://doi.org/10.1145/3506860.3506865

1 INTRODUCTION

Collaborative problem solving refers to the coordinated attempt between two or more people to construct and maintain a shared solution to a problem [35]. A substantial body of research reports numerous benefits of collaborative problem solving such as improving critical thinking [20], developing social skills [37] and learning from diverse viewpoints through constructive argumentation [23]. During the collaborative problem solving process, learners often encounter impasses as part of the natural flow of collaborative interactions. Roschelle et al. [30] referred to an impasse during collaborative problem solving as when team members had either differing opinions regarding the task or insufficient ideas to make progress on the task.

Impasse can be beneficial for learning. VanLehn [38], discussing impasse in the context of individual learning, stated that "Learning occurs only when an impasse occurs. If there is no impasse, there is no learning." D'Mello et al. [9] characterized individual impasse as the learner lacking the knowledge to solve a given task, and suggested that when learners reach an impasse, they need to engage in effortful cognitive activities which can lead to either improved understanding and higher learning gains if the impasse is resolved, or frustration and boredom if the impasse persists. Because working through impasse is important for learning, automatically detecting the moment when learners reach an impasse is a crucial step toward modeling the collaborative problem solving process, as well as informing the development of adaptive learning support for collaboration.

However, there is limited research on impasse in the context of collaborative problem solving, and we have not yet seen techniques for automatically detecting impasse during collaboration. Prior work focused on analyzing individual learners' cognitive and affective transitions when reaching an impasse [9, 26]; however, analyzing individual impasses is insufficient for detecting impasses when learners collaborate in groups. Within collaboration, individual impasse could be resolved when the partner who did not reach the same impasse assists in resolving it [17], but a group impasse indicates that the team as a unit could not move forward because they hold differing opinions or they all individually reached impasse and cannot resolve it without extra help [30]. In this paper, we aim to fill the gap and automate the detection of group impasse during collaborative problem solving, which holds the potential to improve the learning experience by supporting timely interventions when impasse persists.

Our paper explores how multimodal learning analytics can be used to detect impasse during dyadic collaborative problem solving. We collected audio and video data from 46 middle school learners

(23 pairs), who worked on a series of collaborative coding activities. We address the following research questions (RQs):

- RQ1: What are the most predictive unimodal features to detect impasse during dyads' collaborative interactions?
- RQ2: Does multimodal feature fusion help improve impasse detection performance? If so, what are the best multimodal feature combinations from among those we considered?

To answer RQ1, we extracted and examined the performance of a variety of features in linguistic, audio, and video modalities. In the linguistic modality, we evaluated the following: (1) term frequency-inverse document frequency (TF-IDF) [29] features; (2) semantic features generated from Word2Vec [24] and Bidirectional Encoder Representations from Transformers (BERT) [7]; and (3) speaker changes [15]. In the audio modality, we evaluated: (1) acoustic-prosodic features (e.g., loudness, pitch, spectral flux) [25]; and (2) a spectrogram-based feature embedding method generated from VGGish [16]. In the video modality, we evaluated: (1) eye gaze; (2) head movements; and (3) facial Action Units (AUs). The results revealed a series of indicators (e.g., semantics, speaker-changes, pitch variations, and the mutual presence of facial AU23 and AU4) for impasse detection. To answer RQ2, we evaluated different combinations of predictive unimodal features. The results suggest that linguistic + audio + video with early fusion was the best-performing feature combination. This study contributes to the ongoing efforts to use multimodal learning analytics to support collaborative problem solving.

2 RELATED WORK

2.1 Multimodal Learning Analytics in Collaborative Problem Solving

Multimodal learning analytics (MMLA) provides new insights into students' learning through analyzing various streams of data (e.g., speech, faces, gestures) during a learning activity [6]. Previous MMLA research explored data-driven approaches and multimodal modeling in attempts to understand students' learning [2, 22], predict learning performance[8, 31], and construct models of students' interactions [1, 12].

In recent years, there have been increased research efforts toward using MMLA to investigate collaborative problem solving [39]. Grover et al. [14] proposed a framework to capture multimodal data (video, audio, clickstream, and screen capture) from pairs of children as they work together on a pair-programming task, and trained a supervised machine learning model to predict the level of collaboration. Worsley [40] collected gesture, speech, and video data as college students collaborated in pairs to complete engineering design tasks, and utilized MMLA to analyze how these multimodal data relate to students' learning gains. Vrzakova et al. [39] analyzed multimodal data including screen capture, speech, and body movements as triads engaged in a collaborative programming task. Stewart et al. [33] utilized MMLA with dialogues, task contexts, facial expressions, and acoustic-prosodic features generated by triads collaborating to solve a programming task.

Our study differs from these studies in two ways: First, while most aforementioned studies focused on acoustic and visual indicators (e.g., spectral features of audio, body movements) to model learners' interactions, our study also examines linguistic indicators for the task of detecting impasse during collaborative problem solving. Second, we analyze collaborative problem solving using a novel approach that automatically detects moments of impasse through state-of-the-art feature representations generated from pre-trained models.

2.2 Impasse and Learning

Prior research has mostly focused on discussing impasse in the context of individual learning. VanLehn [38] proposed the impasse-driven learning theory, in which learning is facilitated through the resolution of impasses encountered during problem solving. D'Mello et al. [10] suggested that successfully resolving impasse is positively related to learning gains. In another study, D'Mello et al. [9] also proposed a model of affective dynamics in which impasse triggers cognitive disequilibrium (a state of uncertainty), requiring learners to regulate their uncertainty through effortful problem solving in order to restore equilibrium. However, failing to resolve the impasse would lead to frustration and eventually boredom if the impasse persists.

There is limited research on impasse in the context of collaborative learning. Roschelle et al. [30] were among the first to study impasse during collaborative problem solving by analyzing dyads' dialogues. Similarly, Lam [18] analyzed dyads' dialogues when they encountered impasse during collaborative tasks, and found a series of dialogic moves that could be used to identify impasses. There is not yet a technique for automatically detecting impasse during collaboration, and our study aims to fill this gap.

3 DATASET

3.1 Participants and Collaborative Problem Solving Tasks

Our dataset consists of audio/video data from 46 learners (23 pairs) in 7th grade classrooms in a middle (lower secondary) school in the United States [3]. The dataset was collected in 2019 before the COVID-19 pandemic. The learners, 30 girls and 16 boys, collaborated on a series of coding activities, in which they learned fundamental CS concepts such as variables, conditionals, and loops using the Snap! block-based learning environment[1]. The learners followed the pair programming paradigm, in which each dyad shared one computer and switched roles between "driver" and "navigator" during the science-simulation coding activity [4](See Fig. 1).

3.2 Data Collection and Text Transcription

Each collaborative coding activity took around 30 minutes. Dyads were video recorded at 30 fps in 1080p through a front-facing detached camera, and each child wore a lavalier microphone without active noise cancelling capabilities. The audio was recorded by digital sound recorders with a sample rate of 48KHz. After the audio/video data collection process was finished, we used an online transcription service[2] to generate the textual transcript for each dyad. The transcripts included three pieces of information for each spoken utterance: (1) *Starting Time*, in the form of *hour:min:sec*; (2)

[1]https://snap.berkeley.edu
[2]https://www.rev.com

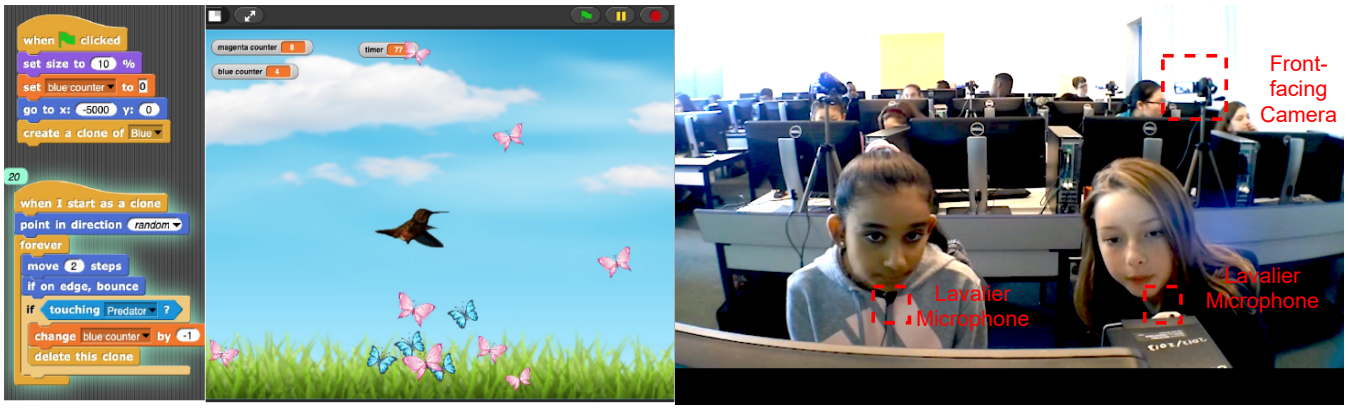

Figure 1: *Left*: A sample script created with Snap!. *Right*: Two middle school learners collaborating on a pair programming task. In the moment, the left learner is the "driver" and the right learner is the "navigator"; their collaborative interaction is video-recorded with a front-facing camera, and audio-recorded with each learner wearing a lavalier microphone.

Speaker, in the form of *S1* or *S2*; and (3) *Transcribed Text*. In total, the corpus included 12 hours and 18 minutes of audio and video recordings, with 10,265 transcribed utterances.

3.3 Impasse Annotation

In line with prior work on manually analyzing impasse during dyadic collaboration [18, 30], our process for annotating impasse was based on the textual transcripts. The tagging protocol followed the idea suggested in prior work [30] that the learners "had differing opinions", or "had insufficient ideas to progress". We annotated impasse at the granularity of one turn exchange, e.g., one back-and-forth conversation of two learners. We referred to these adjacent pairs of turns as "turn exchanges". There were several reasons for this choice of granularity. We opted to annotate turn exchanges from both learners instead of single turns from individuals because the criterion "students had insufficient ideas to progress" requires an indication that both members of the dyad were stuck. At the same time, we did not annotate a longer series of turns because the literature suggests that a next-turn level analysis is a significant indicator within collaborative learning [30], and the sliding window of two turns has substantial computational advantages for reliable detection compared to a longer-size sliding window or a non-fixed-size window.

To prepare for the annotation, two annotators first devised and iteratively refined the annotation protocol on a training set. Then both annotators independently tagged each turn exchange as *Impasse* or *Non-Impasse* in four learning sessions (different than the ones used for training) based on our tagging protocol. The Cohen's Kappa score was 0.70, indicating substantial inter-annotator agreement. Then the first author continued tagging the remaining sessions. Table 1 shows example excerpts and descriptions for both *Impasse* and *Non-Impasse* classes. Among a total of 5,080 turn exchanges, 888 (17.5%) were labeled as *Impasse* and 4,192 (82.5%) were labeled as *Non-Impasse*.

For all turn exchanges, we used FFmpeg[3], an open-source video and audio processing library, to extract the corresponding video and audio segments based on the utterance start time. Since we only obtained the starting time (not end time) of each utterance from the transcription, the video and audio segments included the speech of that turn along with any silence that elapsed before the next turn exchange began. The next step was to extract unimodal features from turn exchange, audio and video segments, and then to train supervised classifiers to identify the most predictive unimodal features for differentiating *Impasse* and *Non-Impasse* samples.

4 FEATURES

4.1 Linguistic Features

Learners often express their disagreements or confusion verbally through dialogue. To extract linguistic features, we represented each turn exchange with two main methods: a statistical method and a semantics-based method.

4.1.1 Statistical Method: Since signal words or phrases (e.g., "*wait*", "*do not*", "*not know*") appear frequently when learners reach an impasse, we experimented with a simple statistical method by generating term frequency-inverse document frequency (**TF-IDF**)[4] ranking of the unigrams and bigrams (consecutive two-word phrases).

4.1.2 Semantics-Based Method: Semantics refers to meaning in language. We utilized three different semantic language models to extract linguistic features: (1) Word2Vec [24], (2) Fine-tuned BERT [7], and (3) Speaker-aware Fine-tuned BERT [15].

(1) Word2Vec: The Word2Vec method learns word associations from the text, and groups similar words together in a vector space based on their semantics. Our Word2Vec model was trained with *gensim*, an open-source natural language processing library[5]. The vector size parameter for each word embedding was set to 300, with a default sliding window size of 5.

(2) Fine-tuned BERT: Similar to Word2Vec, BERT represents semantics of words in a vector space. A pre-trained BERT is a language model that was trained on a large amount of data (e.g., texts

[3]https://www.ffmpeg.org

[4]https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.TfidfVectorizer.html

[5]https://radimrehurek.com/gensim

Table 1: Annotation examples of *Impasse* and *Non-Impasse* during dyads' collaborative dialogues.

Annotation Categories	Turn Exchange Transcripts	Description
Impasse Disagreement	*A: This bottom one? That's right here.* *B: No, it's not.*	Learner A is answering/explaining and learner B disagrees.
	A: Let me see. That started a clone. *B: I don't think that's right.*	Learner A is elaborating and learner B disagrees.
	A: No, when the flag is clicked. *B: No, it says when this costume is clicked.*	Learner A and learner B are arguing and no consensus is reached.
Impasse Insufficient Ideas	*A: Now we are stuck.* *B: Me too. It's supposed to go in here.*	Learner A and learner B both feel stuck.
	A: Wait. Do you wanna... Can we do that? *B: No, yeah, well, I mean, maybe.*	Learner A is asking a question and learner B does not give a certain answer.
	A: Okay. I don't know what to do. *B: I said I don't understand first.*	Learner A and learner B both feel stuck.
Non-Impasse	*A: That's cool.* *B: I'm glad it's moving.*	Learner A and learner B are moving along smoothly with the coding task.
	A: Wait, if the amplitude is... *B: Greater than, yeah, 80.*	Learner A is thinking and learner B gives a certain answer.
	A: Did I create amplitude? No, I didn't. *B: No, you didn't. Not yet.*	Learner A is self-answering and learner B agrees.

from Wikipedia and books) in a self-supervised way. Fine-tuning helps pre-trained BERT models adapt to the specific context of the downstream learning task (in our case, classification of impasse) and achieve better task performance. We fine-tuned the BERT-base-uncased model, which is a publicly available BERT model trained only on English texts[6]. With this model, we transformed each spoken word into a 768-dimensional word embedding vector. The fine-tuning process involves the weights of the hidden layers in the BERT model updated along with the objective of minimizing the loss for the subsequent classification of impasse.

(3) Speaker-aware Fine-tuned BERT: We experimented with a version of fine-tuned BERT that used an additional embedding for speaker information. Speaker information is influential in whether a collaborative dialogue excerpt represents an impasse, as indicated by this concatenated sequence from one turn exchange: *Is this right? I don't know how to get them. Wait, go on this.* Without speaker information, this sequence could have come from different cases of turn exchange:

Case 1:

- *Speaker A: Is this right?*
- *Speaker B: I don't know how to get them. Wait, go on this.*

Case 2:

- *Speaker A: Is this right? I don't know how to get them.*
- *Speaker B: Wait, go on this.*

In the first case, the dyad members experienced a brief period of impasse because learner A was asking a question, but learner B did not give a certain answer. This example belongs to the *Impasse-Insufficient Ideas* category. However, the second case belongs to the *Non-Impasse* category because learner A was asking a question, but received a definite answer from learner B. To differentiate between these two cases, our impasse detection model should be aware of which turn was spoken by which speaker. Following prior work that introduced speaker embeddings into BERT models [15, 36], we first added an *end of turn* segmentation token, [EOT], at the position of the speaker transition in each concatenated sequence, then applied an additional speaker embedding for each word token in the sequence. Following the general pipeline to generate sequence embeddings from BERT [7], the input token representation was the sum of token embeddings, position embeddings, and speaker embeddings. The speaker embeddings were randomly initialized and learned during model training (Fig. 2).

4.2 Acoustic Features

Simple audio indicators (e.g., speaking time, synchrony in the rise and fall of the pitch) derived from audio have been widely used to analyze and assess the quality of collaboration and recognize dyadic affective states [5, 28]. In addition, spectrogram (an image representation that describes an audio's time-frequency distribution) has proven effective in measuring speakers' emotions (e.g., positive, neutral, negative) [27]. Because impasse triggers different affective states in learners[9], we investigated acoustic features for detecting impasse. We used openSMILE, an open-source automatic acoustic feature extraction toolkit [11], for extracting acoustic-prosodic indicators. For each turn exchange's corresponding audio segment, we used openSMILE to extract acoustic-prosodic feature sets (See Table 2) within a 20ms frame and a window shift of 10ms. Apart from acoustic-prosodic features, we also used VGGish[7] for extracting audio embeddings from spectrograms. VGGish is a neural network pre-trained on over 2 million Youtube soundtracks[8] with more than 1,000 human-labeled audio event classes. For each turn exchange's corresponding audio segment, VGGish generated a 128-dimensional vector for every one-second audio frame after dimensionality reduction with Principal Component Analysis.

[6]https://github.com/google-research/bert

[7]https://github.com/tensorflow/models/tree/master/research/audioset/vggish

[8]https://research.google.com/youtube8m

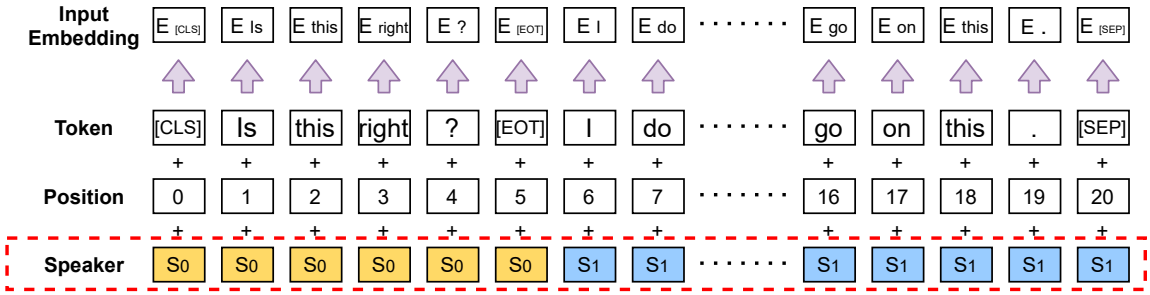

Figure 2: BERT with speaker embedding for Case 1. The input embeddings are the sum of token, position, and speaker embeddings. S_0 and S_1 represent the first and the second speaker of the current turn exchange. [CLS] is the start token aggregating the hidden representation of the whole sequence, and will be used for the later impasse moment classification task; [EOT] is the turn segmentation token indicating the position of speaker switch; [SEP] is the end token of the sequence.

Table 2: Description of the acoustic-prosodic features [25] extracted by openSMILE.

Acoustic-Prosodic Feature	Description
Loudness	A measurement of amplitudes of the signal
Pitch	A measurement of frequencies of the signal
Jitter	How quickly the pitch of the signal is changing
Shimmer	How quickly the loudness of the signal is changing
Spectral Flux	How quickly the power spectrum of the signal is changing
MFCCs	A description of the shape of the signal's short-term spectrum

4.3 Visual Features

A variety of features generated from the video modality have been investigated in prior literature to model collaborative problem solving. Typical features include face tracking [32], facial expressions [33], body movements [39], and distance metrics between learners [6]. Facial behaviors can indicate impasse-related phenomena (e.g. confusion) through movements such as brow lowering and eyelid tightening [13]. In this paper, we used the OpenFace 2.0 facial behavior analysis toolkit[9], which supports accurate facial landmark detection, head pose estimation, eye-gaze direction estimation, and facial Action Unit (AU) recognition for videos containing a single face or multiple faces. We used the *multiple faces* mode to extract three visual features generated from the video modality: eye gaze, head pose, and facial Action Units (AUs).

In each detected face in each video frame, OpenFace generated a 120-dimensional eye gaze vector (112 eye landmarks, 6 eye direction vectors, 2 eye direction in radius), a 6-dimensional head position vector which represents the location of the head with respect to the camera, and a 35-dimensional facial AU vector, including 17 facial AU intensity (0 to 5) features, and 18 facial AU presence (0-absence or 1-presence) features. Facial AUs, which are related to the movements of an individual's facial muscles, have been used in the Facial Action Coding System[10] to taxonomize human facial movements by their appearance on the face. Facial AUs have been used to measure both individual learners' tutoring outcomes [13] and interaction level during collaborative learning [21].

4.4 Feature Aggregation

For linguistic features, word embeddings were concatenated to form the feature vector for each turn exchange. Zero padding[11] was then applied to features of all short turn exchanges to ensure a standard feature dimension for model training. We used a standard word length of 47, which was the maximum length of all turn exchanges in the corpus. For acoustic and visual features, frame-level features were first averaged across a small non-overlapping time window, and then concatenated to form the feature vector for each turn exchange. Following feature aggregation methods in prior work [32, 34], we selected the time window of 500 milliseconds. We did not choose a longer length because acoustic features (e.g, pitch) could vary over a longer time, which would lead to losing fine-grained details. Finally, zero padding was also applied with a standard length of 124, since the maximum elapsed time interval for turn exchanges in the corpus was 62 seconds.

5 EXPERIMENTS AND RESULTS

We built several supervised unimodal and multimodal models trained on the linguistic, acoustic, and visual features to detect *Impasse* moments. Our models (dode available on Github[12]) were implemented in Keras with a Tensorflow backend. We conducted five-fold cross-validation to train and evaluate the models, and the train-test ratio was set to 80% for training and 20% for testing. We evaluated all trained models with F-1 score, combined with precision and recall for both the *Impasse* class and the *Non-Impasse* class.

[9]https://github.com/TadasBaltrusaitis/OpenFace
[10]https://www.cs.cmu.edu/~face/facs.htm
[11]https://www.tensorflow.org/guide/keras/masking_and_padding
[12]https://github.com/yingbo-ma/Detecting-Impasse-LAK2022

Table 3: Testing results for SVM classifiers trained on unimodal features. Label distribution: *Impasse* (17.5%) and *Non-Impasse* (82.5%). Columns: *P* = Precison, *R* = Recall, *F* = F-1 Score, *A* = Accuracy.

		Impasse			*Non-Impasse*			A
Modality	**Unimodal Feature**	**P**	**R**	**F**	**P**	**R**	**F**	
Linguistic	TF-IDF	0.19	0.65	0.30	0.85	0.50	0.64	0.52
	Word2Vec	0.25	0.68	0.37	0.84	0.70	0.76	0.61
	Fine-tuned BERT	0.32	0.70	0.44	0.84	0.77	0.80	0.71
	Fine-tuned BERT + Speaker Embedding	0.36	0.72	**0.48**	0.83	0.76	0.79	**0.76**
Audio	Loudness	0.15	0.23	0.19	0.86	0.64	0.73	0.61
	Pitch	0.19	0.27	0.22	0.86	0.68	0.75	0.64
	Jitter + Shimmer + Spectral Flux	0.17	0.30	0.20	0.85	0.68	0.74	0.62
	MFCCs	0.17	0.56	**0.26**	0.83	0.76	0.79	**0.68**
	VGGish Audio Embedding	0.15	0.62	0.24	0.86	0.43	0.57	0.46
Video	Eye Gaze	0.17	0.63	0.26	0.87	0.34	0.49	0.40
	Head Position	0.18	0.43	0.25	0.85	0.59	0.69	0.56
	Facial AUs	0.20	0.59	**0.31**	0.85	0.68	0.75	**0.66**

5.1 Identifying Predictive Unimodal Features

We trained Support Vector Machine (SVM) classifiers with each of the unimodal features to identify predictive unimodal features for the task of impasse detection. SVMs have shown strong classification performance, especially for small-sized datasets. We used the SVM classifier from scikit-learn[13] with default configurations. Since the label distribution of *Impasse* samples (17.5%) and *Non-Impasse* samples (82.5%) within our corpus is highly imbalanced, we used down-sampling training with the *RandomUnderSampler* from *imblearn*[14]. Without mitigating the effect of the label imbalance, the classifiers would have poor performance for recognizing *Impasse* samples, as they would have been trained mostly on data with *Non-Impasse* samples. Table 3 shows the impasse classification performance trained on unimodal features.

In the linguistic modality, semantic features outperformed statistical features. The Fine-tuned BERT + Speaker Embedding-based classifier yielded the best impasse detection performance, with the highest F-1 score of 0.48 for *Impasse* and the highest overall accuracy of 0.76 for two classes. The TF-IDF-based classifier performed the worst at differentiating impasse and non-impasse, with the lowest accuracy of 0.52. In the audio modality, spectral domain acoustic-prosodic features (Pitch, MFCCs) outperformed time domain features (Loudness), and the MFCCs-based classifier yielded the highest F-1 score of 0.26 for *Impasse* and the highest accuracy of 0.68. The VGGish-based classifier performed worse than classifiers trained on simple acoustic-prosodic features. In the video domain, facial muscle movements outperformed eye gaze directions and head positions, and the Facial AUs-based classifier achieved the highest F-1 score of 0.31 for *Impasse* and the highest accuracy of 0.66. The eye gaze-based classifier performed the worst, with the lowest accuracy of 0.40.

5.2 Examining the Performance of Multimodal Features

Identifying predictive unimodal features helped with filtering out noisy features that potentially had low correlation with impasse detection. Next, we examined the performance of combining these predictive unimodal features into a multimodal model. For the best unimodal features, we selected Fine-tuned BERT + Speaker Embedding from the linguistic modality, MFCCs from audio, and Facial AUs from video, based on the result from Table 3. We experimented with four different combinations of modalities: Linguistic + Audio, Linguistic + Video, Audio + Video, and Linguistic + Audio + Video.

We used a concatenation layer to transform features from different modalities into a single multimodal feature vector. Before concatenating unimodal feature vectors into a single multimodal feature vector, we applied z-score normalization to all the features by subtracting their mean value and dividing by their standard deviation. Next, we provided the concatenated multimodal feature vector as the input to the classifier to perform the binary classification task. We also compared the performance of two different classifiers for this task: SVM and Multi-layer Perceptron (MLP). We chose linear SVM for its strong performance on small-sized datasets and MLP to examine the performance of a non-linear classifier. The MLP classifier contains two feed-forward layers with an embedding size of 128, and two dropout layers with a rate of 0.5 were added to each linear layer respectively to alleviate over-fitting. The *Sigmoid* activation function was used in the last output layer to generate binary classification results. Model weights were updated using an Adam optimizer with the learning rate of $1 \times e^{-5}$. These models were trained for up to 50 epochs, stopping early if validation loss did not decrease for 15 epochs. Fig 3 shows an example multimodal Linguistic + Audio + Video model. The other multimodal models followed the same structure with a subset of the modalities.

Table 4 shows the impasse classification performance of models trained on features combined from different modalities. The results showed that the combination of predictive unimodal features from linguistic, audio and video outperformed all other multimodal combinations for both SVM and MLP classifiers. For the same feature combination, the MLP classifier slightly outperformed the SVM classifier. The multimodal Linguistic + Audio + Video model trained with MLP classifier achieved the best impasse detection performance, with highest F-1 score of 0.65 for *Impasse* and the highest overall accuracy of 0.84 for two classes.

6 DISCUSSION

This paper explores the task of detecting impasse moments within collaborative problem solving activities in middle school classrooms.

[13]https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVC.html
[14]https://pypi.org/project/imbalanced-learn

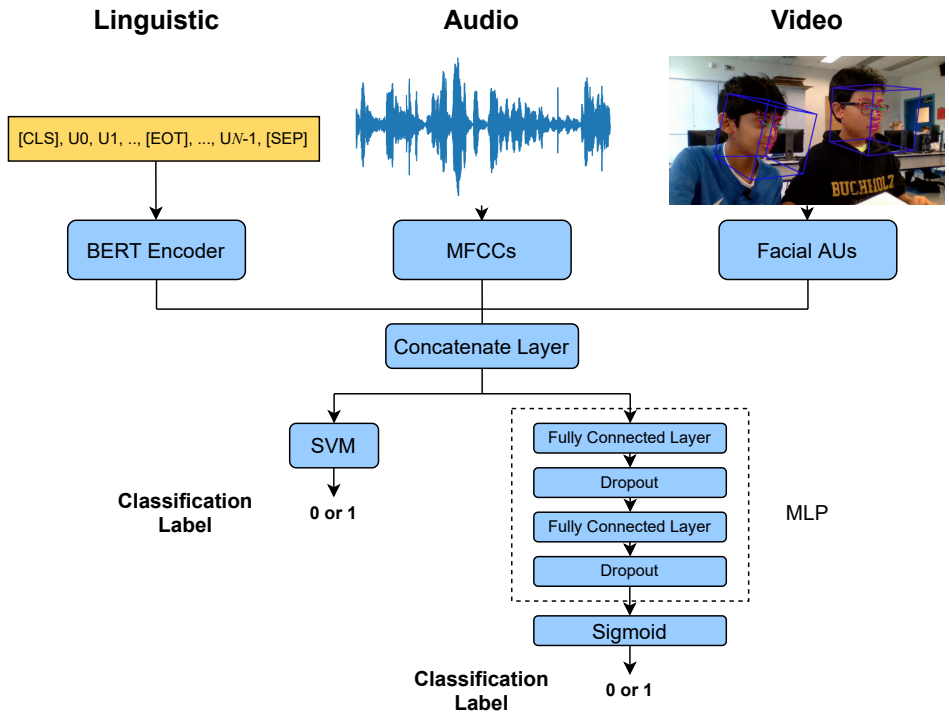

Figure 3: Architecture of the Linguistic + Audio + Video model. The inputs were linguistic, acoustic, and visual features for each turn exchange. The early-fused multimodal feature vector was fed into an SVM or MLP classifier.

Table 4: Results of multimodal combinations of linguistic features (Fine-tuned BERT + Speaker Embedding), acoustic features (MFCCs), and visual features (Facial AUs).

		Impasse			*Non-Impasse*			A
	Modalities Combination	**P**	**R**	**F**	**P**	**R**	**F**	
SVM Classifier	Linguistic + Audio	0.34	0.76	0.48	0.93	0.88	0.90	0.78
	Linguistic + Video	0.43	0.78	0.55	0.94	0.85	0.88	0.80
	Audio + Video	0.25	0.64	0.36	0.85	0.72	0.77	0.69
	Linguistic + Audio + Video	0.55	0.77	**0.63**	0.91	0.90	0.90	**0.81**
MLP Classifier	Linguistic + Audio	0.41	0.79	0.55	0.91	0.90	0.90	0.81
	Linguistic + Video	0.46	0.83	0.60	0.95	0.87	0.91	0.82
	Audio + Video	0.28	0.70	0.39	0.84	0.77	0.80	0.72
	Linguistic + Audio + Video	0.56	0.79	**0.65**	0.96	0.85	0.90	**0.84**

Next, we discuss our results with respect to our two research questions.

6.1 RQ1. Predictive unimodal features to detect impasse during dyadic collaborative interactions.

6.1.1 Linguistic. In the linguistic modality, we experimented with two feature representation methods: the statistical method and the semantics-based method. The results showed that the statistical method (TF-IDF) performed worse than the semantic methods (Word2Vec and the BERT methods). Table 5 shows the most frequently spoken words and phrases in both *Impasse* and *Non-Impasse* samples, with seven out of ten words or phrases spoken in both classes. This suggested that it is insufficient to differentiate impasse and non-impasse by counting the appearance of signal words or phrases that express disagreement or stuck (e.g., "*wait*", "*do not*", "*not know*"). Within the semantic methods, fined-tuned BERT with speaker embedding outperformed Word2Vec and fine-tuned BERT. It is not surprising that the fined-tuned BERT performed better than Word2Vec, since BERT models not only capture a static semantic meaning but also a contextualized meaning compared to Word2Vec. Meanwhile, the additional speaker switch information provided by speaker embedding provides the model with more context to help disambiguate the meanings of utterances. For example, the phrases "*wait*", "*i do not think*" spoken later by the second learner in a turn exchange were usually related to an impasse-disagreement moment, and the phrase "*i do not know*" spoken both by both learners in a

Table 5: Top 10 (from left to right) most frequently spoken unigrams and bigrams in *Impasse* and *Non-Impasse* classes. Unigrams and bigrams are bolded if they appeared in both classes.

					Unigram					
Impasse	***"it"***	***"is"***	***"i"***	*"wait"*	***"do"***	***"to"***	***"the"***	*"not"*	*"no"*	***"we"***
Non-impasse	***"it"***	***"is"***	***"the"***	***"i"***	***"to"***	*"okay"*	***"we"***	*"that"*	*"oh"*	***"do"***
					Bigram					
Impasse	***"it is"***	***"do not"***	***"i do"***	*"not know"*	*"wait wait"*	***"i think"***	*"no no"*	***"i am"***	***"that is"***	***"have to"***
Non-impasse	***"it is"***	*"and then"*	***"do not"***	***"i am"***	***"that is"***	***"have to"***	***"i do"***	***"i think"***	*"to do"*	*"we have"*

Table 6: Top 5 (from left to right) detected facial AUs from both two learners in a dyad.

Impasse	AU 1 Inner Brow Raiser	AU 2 Outer Brow Raiser	AU 23 Lip Tightener	AU 15 Lip Corner Depressor	AU 4 Brow Lowerer
Non-Impasse	AU 1 Inner Brow Raiser	AU 2 Outer Brow Raiser	AU 20 Lip Stretcher	AU 15 Lip Corner Depressor	AU 25 Lips Part

turn exchange was usually related to an impasse-no-sufficient-ideas moment.

6.1.2 Audio. In the audio modality, we compared a set of acoustic-prosodic features and the VGGish audio spectrogram embedding method. The VGGish audio spectrogram embedding performed worse than acoustic-prosodic features. Within the set of acoustic-prosodic features, spectral domain features (e.g. Pitch, MFCCs) generally outperformed time domain features (e.g. Loudness, Shimmer). Figure 4 compares the the pitch and loudness (plotted by Praat[15]) of an *Impasse-Disagreement* and a *Non-Impasse* audio segment. As illustrated on the impasse plot, when learner B started saying "*No, it's not supposed...*", the pitch significantly rose while the loudness was steady compared to previous moments when learner A was talking. The non-impasse plot shows that when learner B started saying "*No, you didn't...*", neither pitch nor loudness rose significantly compared to previous moments when learner A was talking. This phenomenon potentially suggests that measuring the pitch variance when the second learner started talking could be an effective way to detect impasse moments when learners engage in argument and cannot reach an consensus. Our finding is aligned with prior work [19] in which the authors observed that the synchrony in the rise and fall of the pitch between two learners was the most significant acoustic-prosodic feature of when rapport was present in collaborative dialogues.

6.1.3 Video. In the video modality, we experimented with eye gaze, head position, and facial AUs, all of which were automatically generated with the OpenFace 2.0 facial behavior analysis toolkit. The results showed that the presence and intensity of facial AUs outperformed eye gaze and head position. Specifically, we observed that the mutual presence of some common facial AUs from both learners could be indicative a dyad's impasse moments. Table 6 shows the most frequently present facial AUs detected from both learners. As illustrated on the table, the most distinctive facial AU patterns for detecting impasse were AU 23, Lip Tightener, and AU 4, Brow Lowerer. AU 20, Lip Stretcher, and AU 25, Lips Part, were most predictive for non-impasse.

6.2 RQ2. Multimodal feature fusion for detecting impasse during dyadic collaborative dialogue.

A main thrust of our work was to automatically detect impasse moments with supervised learning models and to identify the best multimodal feature combinations from among those we considered. After selecting the best-performing unimodal features within each modality, we experimented with four different modality combinations. The results showed that Linguistic + Audio + Video (F1 Score = 0.65) yielded the best impasse detection performance, and Audio + Video (F1 Score = 0.39) performed the worst. There are several important implications behind these results. First, all of our multimodal models outperformed their unimodal models (e.g., the Audio + Video model outperformed unimodal Audio / Video models, the Linguistic + Audio model outperformed unimodal Linguistic / Audio models), which indicated that combining modalities improves impasse detection performance. Second, the close performance from the Linguistic + Video model and the Linguistic + Audio + Video model suggested that adding acoustic-prosodic features to the Linguistic + Video model did not make a substantial difference. As we discussed in section 6.1.2, the significant rise in pitch when the second learner started talking was more correlated with impasse moments when learners expressed disagreement; however, this pattern may not be observed with impasse moments when learners are both stuck and do not have sufficient ideas to proceed. The noisy classroom environments may be another potential reason for the poor impasse detection performance of acoustic-prosodic features: Our corpus was collected with middle school students collaborating face-to-face in their classrooms; when dyads were engaging in collaborative problem solving tasks, the background speech from other dyads was audible. Sudden high-pitched background noises (e.g., a chair moving) were also an issue. In these noisy cases, acoustic features (and linguistic features if the texts were transcribed with automatic speech-to-text services) would not be as reliable as visual features. Fig 5 (shown on the next page) depicts three samples (two impasse and one non-impasse) and their corresponding classification results from different modality combinations.

[15]https://www.fon.hum.uva.nl/praat

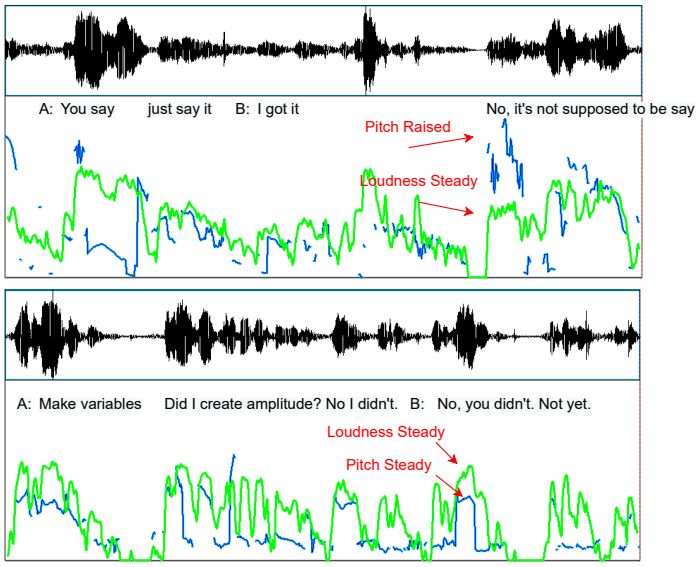

Figure 4: *Top: Impasse-Disagreement.* The pitch (blue) rose significantly when learner B started talking while the loudness (green) was steady. *Bottom: Non-Impasse.* The pitch (blue) and loudness level (green) were steady when learner B started talking.

6.3 Limitations.

The current work has several important limitations. First, our annotation was based on textual transcripts, and the video and audio segments of our corpus included the speech of each turn along with any silence that elapsed before the next turn exchange started. Therefore, we could not explicitly model silence and pauses during dyads' collaborative dialogue, and these phenomena may hold useful information that can help to more accurately detect impasse: for example, a long silence during a dyad's interaction may suggest learners being stuck. Second, there was one ubiquitous limitation in facial detection work: OpenFace sometimes failed to detect both learners' faces when they were not directly facing the camera, or in the case of occlusion. Finally, our corpus size was relatively small. The recordings were collected from just 46 middle school learners; therefore, the predictive unimodal features found in this paper may not generalize well to students in other age groups or learning environments, such as online learning.

7 CONCLUSION AND FUTURE WORK

Impasse occurs when learners cannot not move forward because they have differing opinions or insufficient ideas. While impasse presents important opportunities for learning, it can also negatively impact problem solving when it persists for too long. By automatically detecting impasse moments during collaborative problem solving, we can support the development of systems that better support learners. This paper has presented the first attempt to automatically detect dyadic impasse moments during collaborative problem solving. We combined linguistic features from BERT, acoustic-prosodic features from openSMILE, and visual features from the OpenFace facial behavior analysis toolkit. The results revealed a series of important verbal and non-verbal indicators for detecting impasse, and the multimodal Linguistic + Audio + Video model performed the best for this task.

These results highlight several important directions for future work. First, while the features used in this paper were promising, other features should be investigated (e.g. silence and pauses during speech, body movements). It is also necessary to investigate features that are less hand-crafted as we are moving toward detecting impasse in real-time. Second, the multimodal model's performance could potentially be increased with more optimal classifier configurations or effective feature extraction methods (e.g., representing speaker change with pitch variations). Third, future work should examine generalizability of the findings in this work using larger datasets, including data from online learning environments. To create an impasse detector that could be used in many different learning scenarios, it will also be important to determine how the features of impasse moments differ with learners of different ages and cultures. Finally, we aim to investigate impasse detection in multi-party interactions among groups of three or more learners.

Impasse moments provide valuable and beneficial learning opportunities for team members during their collaborative problem solving process. The work presented in this paper makes a step toward automatically detecting the impasse resolution process in real-time. This line of investigation has the potential to improve the learning experience by supporting timely interventions when impasse persists.

ACKNOWLEDGMENTS

This research was supported by the National Science Foundation through grant DRL-1640141. Any opinions, findings, conclusions, or recommendations expressed in this report are those of the authors, and do not necessarily represent the official views, opinions, or policy of the National Science Foundation.

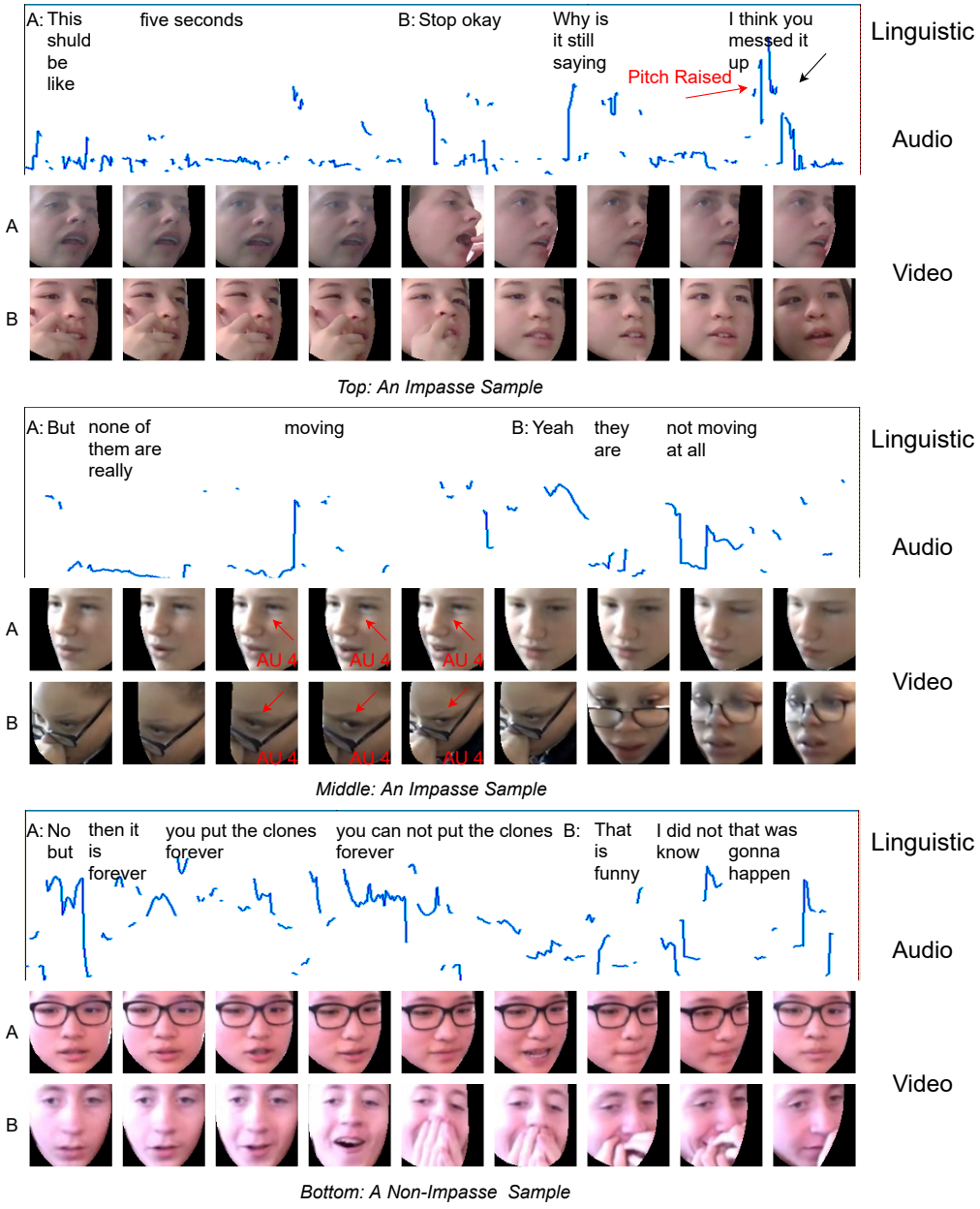

Figure 5: *Top*: This *Impasse* sample has acoustic indicators (pitch raised) for impasse. It was correctly classified by Linguistic + Audio model and Linguistic + Audio + Video model, but misclassified by Linguistic + Video model. *Middle*: This *Impasse* sample has visual indicators (co-present AU 4 from two learners' faces) for impasse. It was correctly classified by the Linguistic + Video and Linguistic + Audio + Video models, but misclassified by the Linguistic + Audio model. *Bottom*: This *Non-Impasse* sample has neither acoustic indicators nor visual indicators for impasse. It was was correctly classified by the Audio + Video and Linguistic + Audio + Video models, but misclassified by the Linguistic + Audio and Linguistic + Video models.

REFERENCES

[1] Laura K Allen, Caitlin Mills, Matthew E Jacovina, Scott Crossley, Sidney D'mello, and Danielle S McNamara. 2016. Investigating boredom and engagement during writing using multiple sources of information: the essay, the writer, and keystrokes. In *Proceedings of the Sixth International Conference on Learning Analytics & Knowledge*. 114–123.

[2] Alejandro Andrade. 2017. Understanding student learning trajectories using multimodal learning analytics within an embodied-interaction learning environment. In *Proceedings of the Seventh International Conference on Learning Analytics & Knowledge*. 70–79.

[3] Mehmet Celepkolu, David Austin Fussell, Aisha Chung Galdo, Kristy Elizabeth Boyer, Eric N Wiebe, Bradford W Mott, and James C Lester. 2020. Exploring middle school students' reflections on the infusion of CS into science classrooms. In *Proceedings of the 51st ACM technical symposium on computer science education*. 671–677.

[4] Mehmet Celepkolu, Aisha Chung Galdo, Kristy Elizabeth Boyer, Eric N Wiebe, Bradford Mott, and James C Lester. 2020. Student Reflections on Pair Programming in Middle School: A Thematic Analysis. (2020).

[5] Huili Chen, Yue Zhang, Felix Weninger, Rosalind Picard, Cynthia Breazeal, and Hae Won Park. 2020. Dyadic speech-based affect recognition using DAMI-P2C parent-child multimodal interaction dataset. In *Proceedings of the International Conference on Multimodal Interaction*. 97–106.

[6] Mutlu Cukurova, Qi Zhou, Daniel Spikol, and Lorenzo Landolfi. 2020. Modelling collaborative problem-solving competence with transparent learning analytics: is video data enough?. In *Proceedings of the Tenth International Conference on Learning Analytics & Knowledge*. 270–275.

[7] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).

[8] Daniele Di Mitri, Maren Scheffel, Hendrik Drachsler, Dirk Börner, Stefaan Ternier, and Marcus Specht. 2017. Learning pulse: A machine learning approach for predicting performance in self-regulated learning using multimodal data. In *Proceedings of the Seventh International Conference on Learning Analytics & Knowledge*. 188–197.

[9] Sidney D'Mello and Art Graesser. 2012. Dynamics of affective states during complex learning. *Learning and Instruction* 22, 2 (2012), 145–157.

[10] Sidney D'Mello, Blair Lehman, Reinhard Pekrun, and Art Graesser. 2014. Confusion can be beneficial for learning. *Learning and Instruction* 29 (2014), 153–170.

[11] Florian Eyben, Martin Wöllmer, and Björn Schuller. 2010. openSMILE: the munich versatile and fast open-source audio feature extractor. In *Proceedings of the International Conference on Multimedia*. 1459–1462.

[12] Aysu Ezen-Can, Joseph F Grafsgaard, James C Lester, and Kristy Elizabeth Boyer. 2015. Classifying student dialogue acts with multimodal learning analytics. In *Proceedings of the Fifth International Conference on Learning Analytics & Knowledge*. 280–289.

[13] Joseph F Grafsgaard, Joseph B Wiggins, Kristy Elizabeth Boyer, Eric N Wiebe, and James C Lester. 2013. Automatically recognizing facial indicators of frustration: a learning-centric analysis. In *2013 Humaine Association Conference on Affective Computing and Intelligent Interaction*. IEEE, 159–165.

[14] Shuchi Grover, Marie Bienkowski, Amir Tamrakar, Behjat Siddiquie, David Salter, and Ajay Divakaran. 2016. Multimodal analytics to study collaborative problem solving in pair programming. In *Proceedings of the Sixth International Conference on Learning Analytics & Knowledge*. 516–517.

[15] Jia-Chen Gu, Tianda Li, Quan Liu, Zhen-Hua Ling, Zhiming Su, Si Wei, and Xiaodan Zhu. 2020. Speaker-aware BERT for multi-turn response selection in retrieval-based chatbots. In *Proceedings of the Tenth International Conference on Information & Knowledge Management*. 2041–2044.

[16] Shawn Hershey, Sourish Chaudhuri, Daniel PW Ellis, Jort F Gemmeke, Aren Jansen, R Channing Moore, Manoj Plakal, Devin Platt, Rif A Saurous, Bryan Seybold, et al. 2017. CNN architectures for large-scale audio classification. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 131–135.

[17] Michelle E Jordan and Reuben R McDaniel Jr. 2014. Managing uncertainty during collaborative problem solving in elementary school teams: The role of peer influence in robotics engineering activity. *Journal of the Learning Sciences* 23, 4 (2014), 490–536.

[18] Rachel Lam. 2019. What students do when encountering failure in collaborative tasks. *NPJ science of learning* 4, 1 (2019), 1–11.

[19] Nichola Lubold and Heather Pon-Barry. 2014. Acoustic-prosodic entrainment and rapport in collaborative learning dialogues. In *Proceedings of the 2014 ACM workshop on Multimodal Learning Analytics Workshop and Grand Challenge*. 5–12.

[20] Louis Major, Paul Warwick, Ingvill Rasmussen, Spliid Ludvigsen, and Victoria Cook. 2018. Classroom dialogue and digital technologies: A scoping review. *Education and Information Technologies* 23, 5 (2018), 1995–2028.

[21] Jonna Malmberg, Sanna Järvelä, Jukka Holappa, Eetu Haataja, Xiaohua Huang, and Antti Siipo. 2019. Going beyond what is visible: What multichannel data can reveal about interaction in the context of collaborative learning? *Computers in Human Behavior* 96 (2019), 235–245.

[22] Roberto Martinez-Maldonado, Vanessa Echeverria, Olga C Santos, Augusto Dias Pereira Dos Santos, and Kalina Yacef. 2018. Physical learning analytics: A multimodal perspective. In *Proceedings of the Eighth International Conference on Learning Analytics & Knowledge*. 375–379.

[23] Neil Mercer, Sara Hennessy, and Paul Warwick. 2017. Dialogue, thinking together and digital technology in the classroom: Some educational implications of a continuing line of inquiry. *International Journal of Educational Research* 97 (2017), 187–199.

[24] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781* (2013).

[25] Puneet Kumar Mongia and RK Sharma. 2014. Estimation and statistical analysis of human voice parameters to investigate the influence of psychological stress and to determine the vocal tract transfer function of an individual. *Journal of Computer Networks and Communications* 2014 (2014).

[26] Brendan Munzar, Krista R Muis, Courtney A Denton, and Kelsey Losenno. 2021. Elementary students' cognitive and affective responses to impasses during mathematics problem solving. *Journal of Educational Psychology* 113, 1 (2021), 104.

[27] Sandra Ottl, Shahin Amiriparian, Maurice Gerczuk, Vincent Karas, and Björn Schuller. 2020. Group-level Speech Emotion Recognition Utilising Deep Spectrum Features. In *Proceedings of the International Conference on Multimodal Interaction*. 821–826.

[28] Sambit Praharaj, Maren Scheffel, Marcel Schmitz, Marcus Specht, and Hendrik Drachsler. 2021. Towards automatic collaboration analytics for group speech data using learning analytics. *Sensors* 21, 9 (2021), 3156.

[29] Juan Ramos. 2003. Using TF-IDF to determine word relevance in document queries. In *Proceedings of the First Instructional Conference on Machine Learning*, Vol. 242. Citeseer, 29–48.

[30] Jeremy Roschelle and Stephanie D Teasley. 1995. The construction of shared knowledge in collaborative problem solving. In *Computer Supported Collaborative Learning*. Springer, 69–97.

[31] Kshitij Sharma, Zacharoula Papamitsiou, Jennifer K Olsen, and Michail Giannakos. 2020. Predicting learners' effortful behaviour in adaptive assessment using multimodal data. In *Proceedings of the Tenth International Conference on Learning Analytics & Knowledge*. 480–489.

[32] Daniel Spikol, Emanuele Ruffaldi, Lorenzo Landolfi, and Mutlu Cukurova. 2017. Estimation of success in collaborative learning based on multimodal learning analytics features. In *2017 IEEE 17th International Conference on Advanced Learning Technologies (ICALT)*. IEEE, 269–273.

[33] Angela EB Stewart, Zachary Keirn, and Sidney K D'Mello. 2021. Multimodal modeling of collaborative problem-solving facets in triads. *User Modeling and User-Adapted Interaction* (2021), 1–39.

[34] Angela EB Stewart, Zachary A Keirn, and Sidney K D'Mello. 2018. Multimodal modeling of coordination and coregulation patterns in speech rate during triadic collaborative problem solving. In *Proceedings of the 20th ACM International Conference on Multimodal Interaction*. 21–30.

[35] Chen Sun, Valerie J Shute, Angela Stewart, Jade Yonehiro, Nicholas Duran, and Sidney D'Mello. 2020. Towards a generalized competency model of collaborative problem solving. *Computers & Education* 143 (2020), 103672.

[36] Anna Tigunova, Paramita Mirza, Andrew Yates, and Gerhard Weikum. 2021. PRIDE: Predicting Relationships in Conversations. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. 4636–4650.

[37] Julianne C Turner, Andrea Christensen, Hayal Z Kackar-Cam, Meg Trucano, and Sara M Fulmer. 2014. Enhancing Students' Engagement: Report of a 3-Year Intervention With Middle School Teachers. *American Educational Research Journal* 51, 6 (2014), 1195–1226.

[38] Kurt VanLehn. 1988. Toward a theory of impasse-driven learning. In *Learning Issues for Intelligent Tutoring Systems*. Springer, 19–41.

[39] Hana Vrzakova, Mary Jean Amon, Angela Stewart, Nicholas D Duran, and Sidney K D'Mello. 2020. Focused or stuck together: Multimodal patterns reveal triads' performance in collaborative problem solving. In *Proceedings of the Tenth International Conference on Learning Analytics & Knowledge*. 295–304.

[40] Marcelo Worsley. 2018. (Dis) engagement matters: Identifying efficacious learning practices with multimodal learning analytics. In *Proceedings of the Eighth International Conference on Learning Analytics & Knowledge*. 365–369.