How Noisy is Too Noisy? The Impact of Data Noise on Multimodal Recognition of Confusion and Conflict During Collaborative Learning

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

Eric Wiebe
North Carolina State University
Raleigh, USA
wiebe@ncsu.edu

Collin F. Lynch
North Carolina State University
Raleigh, USA
cflynch@ncsu.edu

Maya Israel
University of Florida
Gainesville, USA
misrael@coe.ufl.edu

ABSTRACT

Intelligent systems to support collaborative learning rely on real-time behavioral data, including language, audio, and video. However, noisy data, such as word errors in speech recognition, audio static or background noise, and facial mistracking in video, often limit the utility of multimodal data. It is an open question of how we can build reliable multimodal models in the face of substantial data noise. In this paper, we investigate the impact of data noise on the recognition of confusion and conflict moments during collaborative programming sessions by 25 dyads of elementary school learners. We measure language errors with word error rate (*WER*), audio noise with speech-to-noise ratio (*SNR*), and video errors with frame-by-frame facial tracking accuracy. The results showed that the model's accuracy for detecting confusion and conflict in the language modality decreased drastically from 0.84 to 0.73 when the *WER* exceeded 20%. Similarly, in the audio modality, the model's accuracy decreased sharply from 0.79 to 0.61 when the *SNR* dropped below 5 dB. Conversely, the model's accuracy remained relatively constant in the video modality at a comparable level (> 0.70) so long as at least one learner's face was successfully tracked. Moreover, we trained several multimodal models and found that integrating multimodal data could effectively offset the negative effect of noise in unimodal data, ultimately leading to improved accuracy in recognizing confusion and conflict. These findings have practical implications for the future deployment of intelligent systems that support collaborative learning in actual classroom settings.

CCS CONCEPTS

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

KEYWORDS

Data Noise; Multimodal Fusion; Collaborative Learning

ACM Reference Format:
Yingbo Ma, Mehmet Celepkolu, Kristy Elizabeth Boyer, Eric Wiebe, Collin F. Lynch, and Maya Israel. 2023. How Noisy is Too Noisy? The Impact of Data Noise on Multimodal Recognition of Confusion and Conflict During Collaborative Learning. In *INTERNATIONAL CONFERENCE ON MULTIMODAL INTERACTION (ICMI '23), October 9–13, 2023, Paris, France.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3577190.3614127

ICMI '23, October 9–13, 2023, Paris, France

ACM ISBN 979-8-4007-0055-2/23/10…$15.00
https://doi.org/10.1145/3577190.3614127

1 INTRODUCTION

Collaborative learning refers to two or more learners working together to solve a problem, complete a task, or create a shared product [23]. During collaborative activities, learners face cognitive and social challenges and require timely support [34]. One solution for providing such support is to develop intelligent systems for collaborative learning, which can scaffold productive learning and support effective collaboration among learners [27]. These systems rely on accurate modeling of learners' collaborative interactions using their behavioral data from multiple modalities, such as language, audio, and video [48]. Each modality provides unique insights into the collaborative learning process, and combining them may lead to improved accuracy in modeling collaborative learning [33].

In a traditional classroom setting, learners engage in conversation via verbal communication (e.g., verbal affirmations) and through para-verbal cues (e.g., change in speech prosody, facial expressions) [5]. However, noise that often accompanies these behavioral data poses a great challenge for intelligent systems to analyze and understand learners' dialogues [22, 37]. For example, it is difficult to separate learners' speech from background audio noise given the presence of ambient sound or chatter in the surroundings [26]. Furthermore, even when learners' speech can be successfully separated from background audio noise, language errors can still arise because of the imperfect speech recognition from dropped audio, misunderstanding of words due to slang, dialect, as well as young learners' distinct vocal characteristics [4, 47]. In addition, learners' faces may be mistracked when they are not directly facing the camera or in the case of occlusion [15, 51]. When intelligent systems integrate these noisy data, they may not accurately model

the complex dynamics of learners' collaboration processes. By understanding the extent to which these data noise affect modeling accuracy, we can obtain valuable insights for implementing intelligent systems that model collaborative behavior in actual classroom settings.

This paper takes a first step toward addressing the challenge of modeling collaborative learning behavior from error-prone data streams by investigating the impact of language error, audio noise, and facial mistracking on detecting two important moments during collaborative learning: confusion and conflict. We chose to model confusion and conflict because they represent critical moments when learners face cognitive and social challenges during collaborative learning. *Confusion* is an important cognitive-affective state that may emerge when learners face cognitive challenges during collaborative learning [39]; *conflict* is another which captures a state of disagreement or opposition between two or more learners [1]. By automatically detecting conflict and confusion during collaboration, intelligent systems can offer timely assistance to help learners work through the confusion and conflict and improve their learning experience.

Our dataset includes the audio and video recordings of 25 paired elementary school learners working on a series of coding tasks. Specifically, we investigate two research questions (RQs):

- **RQ 1**: To what extent do language errors, audio noise, and facial mistracking impact the accuracy of modeling confusion and conflict during collaborative learning?
- **RQ 2**: To what extent does integrating language, audio, and facial behaviors compensate for the negative impact of unimodal data noise?

We measure language errors with word error rate (*WER*) [6], audio noise with posterior speech-to-noise ratio (*SNR*) [50], and video errors with the number of mistracked faces. To answer RQ1, we compared the performance of text language, audio, and facial behaviors with the error metrics described above. We found a *WER* threshold of around 20% in text language, where the model accuracy for detecting confusion and conflict started to decrease drastically (overall accuracy from 0.84 to 0.73), with F-1 scores dropping from 0.55 to 0.40 for confusion and dropping from 0.53 to 0.37 for conflict. For audio, the model's accuracy decreased sharply from 0.79 to 0.61 when the signal-to-noise ratio (*SNR*) dropped below 5 dB, with F-1 scores dropping from 0.55 to 0.40 for confusion and dropping from 0.53 to 0.37 for conflict. For facial behaviors, compared to when both learners' faces were tracked, where the accuracy was 0.79, the model's accuracy slightly decreased to 0.71 when only the listener's face was tracked, with F-1 scores of 0.51 to 0.44 for confusion and 0.41 to 0.33 for conflict. To answer RQ2, we compared the performance of several multimodal models trained with model-agnostic and model-based fusion methods, and we found that combining multimodal data from language, audio, and facial features can effectively compensate for the negative impact of unimodal data noise, ultimately leading to improved accuracy in recognizing confusion and conflict.

To the best of our knowledge, this work is the first to investigate the impact of data noise on the multimodal modeling of collaborative dialogues. The contributions of this paper are twofold: First, we provide extensive experimental results demonstrating the impact of language errors, audio noise, and facial mistracking errors on modeling collaborative dialogues. Second, we provide empirical evidence for the combination of multimodal data as an effective means of compensating for the negative impact of data noise present in a single modality. These findings provide valuable insights for the future implementation of these models in actual classroom settings.

2 RELATED WORK

2.1 Data Noise

Noisy data may occur due to dynamic environmental conditions, faulty detectors, or other unavoidable quality degradation in the measurements [14]. The impact of data noise has been studied in various domains, such as audio-visual speech recognition [37], risk prediction [16], object tracking [17], and medical diagnosis [61]. For example, in the domain of audio-visual speech recognition, Papandreou et al. [37] studied the impact of data noise under challenging conditions, such as when the visual front-end momentarily mistracks the speaker's face or in noisy acoustic environments. In the field of risk prediction, Heo et al. [16] investigated the effects of data noise on the clinical risk prediction of electronic health records, which frequently suffer from varying degrees of noise and missing entry problems. Medical diagnosis is yet another domain that is prone to data noise, as highlighted by Reyes-Garcia et al. [41], who evaluated the impact of missing values of vital biometrics, such as arterial blood pressure and heart rate, on the early prediction of patients' physiological deterioration. The results of these studies suggest that data noise can negatively impact data analysis and lead to incorrect conclusions, unreliable predictions, or flawed decision-making [32], and it is important to develop effective methods to handle data noise to ensure the accuracy and reliability of the results obtained.

2.2 Multimodal Modeling of Collaborative Learning

Prior research on multimodal modeling of collaborative learning has analyzed learners' interactions across multiple modalities, including speech, facial expressions, body gestures, and physiological data [9, 25, 49]. Analyzing collaborative learning processes by combining multiple modalities of data has shown great promise in building more accurate models. For example, Olsen et al. [35] combined learners' audio, eye gaze, and tutor logs to predict collaborative learning outcomes. Vrzakova et al. [54] analyzed multimodal data, including screen capture, speech, and body movements as triads engaged in a collaborative programming task. Moulder et al. [30] modeled how students' multimodal dynamics (e.g., emotional, verbal communication, physiological) were influenced by each other while engaged in collaborative problem-solving. Eloy et al. [10] used speech rate, body movement, and galvanic skin response to model triads' emotional valence and task performance while collaborating on solving physics games. Although the above-mentioned literature has highlighted the improved accuracy of multimodal modeling over unimodal modeling, the impact of noise and missing values in different modalities, and their collective impact when multimodal data is integrated, has not been investigated in research on modeling collaborative learning, and our study aims to fill this gap.

3 DATASET

3.1 Participants and Collaborative Learning Tasks

Our dataset consists of audio and video data from 25 pairs in fourth-grade classrooms in an elementary school in the southeastern United States. The dataset was collected in the spring of 2022. These learners had an average age of 10, with 21 of them reporting their gender as girls, 12 as boys, and five preferring not to answer. Learners collaborated on a series of coding activities in which they learned fundamental CS concepts such as variables, conditionals, and loops using a block-based learning environment built upon *Snap!* [46]. The learners followed the *pair programming* paradigm, in which each pair (or dyad) shared one computer and switched roles between *"driver"* and *"navigator"* during the science-simulation coding activity. The *driver* is responsible for writing the code and implementing the solution, while the *navigator* provides support by catching mistakes and providing feedback on the in-progress solution [7] (See Fig. 1-Top).

Figure 1: Top: Block-based coding tasks. Bottom: A dyad of learners collaborating together

3.2 Data Collection and Transcription

Each collaborative coding activity took around 40 minutes. Dyads were video-recorded by the front-facing camera of their laptop and audio-recorded with each learner wearing a headset without any additive noise cancellation equipment (See Fig. 1-Bottom). After the data collection, we used an online transcription service [40] to manually 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) *Speaker*, in the form of *S1* or *S2*; and (3) *Transcribed Text*. In total, the corpus included 22 hours and 18 minutes of audio and video recordings, with 9,943 transcribed utterances. We used the timestamp from each spoken utterance to segment the audio and video recordings, generating an audio and corresponding video clip of each spoken utterance.

3.3 Manual Annotation of Confusion and Conflict Dialogues

In line with prior work on analyzing confusion and conflict dialogues during collaborative learning [43, 53], the process of annotating confusion and conflict was based on textual transcripts with video used in rare cases of unresolvable ambiguity within the transcripts. The dialogue act taxonomy draws upon a closely related dialogue act taxonomy by Zakaria et al. [60] that was designed for elementary school learners' classroom dialogues. Table 1 shows example excerpts and descriptions for confusion and conflict.

To establish the reliability of the dialogue act labeling, two annotators first engaged in a training phase where they collaboratively applied the dialogue act taxonomy and discussed any disagreements. Once training was complete, they independently tagged an overlapping 20% of the data, reaching a Cohen's kappa score of 0.816, indicating a strong agreement. They then proceeded to divide and tag the remaining data independently. Among a total of 9,943 transcribed utterances, 467 (4.7%) were labeled as confusion, 924 (9.3%) as conflict, and 8,552 (86.0%) other.

4 DATA NOISE MEASUREMENT

4.1 Errors in Text Language

We measured noise in language by word error rate (WER)[44]. WER is given by $WER = (S + D + I)/N$, where S is the number of substitutions, D is the number of deletions, I is the number of

Table 1: Annotation examples of confusion and conflict dialogues

Category	Example Transcripts	Description	Count (Percentage)
Confusion	I have no idea what to do. I don't know why it's doing that. I'm confused and I don't understand.	Learner is directly or indirectly seeking help from a partner.	467 (4.7%)
Conflict	You are being ridiculous. Well, I don't think so, and that is wrong. No, because that won't make it move in a square.	Actions or interactions that cause tension. Disagreement on any opinions/code editings. Learner disagrees but then explains why.	924 (9.3%)
Other	What does that block do? That looks good. How about doubling that? Thanks, we know we are great. Give me the keyboard.	Learner asks questions. Agreement on any opinions/code editings. Suggestions directly talking to a partner. Social dialogues. Directive, telling partners to do something.	8,552 (86.0%)

insertions, and N is the number of words in the human transcript. For each human-transcribed utterance, additive noise was manually generated by randomly substituting, deleting, or inserting words, to create five different noisy transcripts with a WER of 0.1, 0.2, 0.3, 0.4, and 0.5. In this study, we initially tested several transcription engines in an effort to obtain WERs low enough for our study. We experimented with both commercial engines and open-source engines (e.g., Google Speech-to-Text [52] and OpenAI Whisper [58]), and these models all generated overall high WERs (above 0.70). Consequently, in order to investigate data with lower WER, we decided to manually introduce a controlled amount of textual noise. The manual perturbation steps for each utterance are: (1) randomly select an action (i.e., *substitution*, *deletion*, or *insertion*), (2) randomly select a word within the given utterance, (3) if the action is *deletion*, delete the word selected from the last step; if the action is *substitution*, randomly select another word in the given utterance, then substitute the word with the word selected from the last step; if the action is *insertion*, randomly select a position in the given utterance, then insert the word selected from the last step to the selected position.

4.2 Noise in Audio

We measured noise in the audio modality by using the posterior signal-to-noise ratio (SNR) [50], which is the ratio of signal power to noise power, often expressed in decibels (dBs). An SNR value of 0 dB indicates the same power of signal and noise. We calculated SNR by $SNR_{post}(t) = log\frac{E(t)}{E_{noise}(t)}$, where $E(t)$ is the energy of noisy speech of audio frame t, and $E_{noise}(t)$ is the energy of noise of frame t. We estimated $E_{noise}(t)$ by averaging the $E_{noise}(t)$ of each silent audio frame (identified by Silero [45], a pre-trained voice activity detection model), then used the averaged $E_{noise}(t)$ to calculate $SNR_{post}(t)$ for each speech audio segment (an SNR of +15 dB or above indicates *good* speech quality). The overall average SNR of our dataset is +1.3 dB. Table 2 shows the number of different audio segments in the dataset that fall into different classes.

Table 2: Number of audio segments of each SNR level

SNR level (dB)	Confusion	Conflict	Other
5 or greater	93	211	1,759
0 to 5	186	347	2,804
-5 to 0	139	278	2,365
less than -5	49	88	624
Total	467	924	8,552

4.3 Mistracked Faces in Video

We partitioned the face recognition data into segments and classified the segments into one of four cases: **(1) Both learners' faces were tracked.** This is the optimal condition; at the beginning of each dyad's learning session, their laptop was set in the middle of them to track both their faces. **(2) Only the speaker's face was tracked.** This error may happen when learners adjusted the direction of the laptop or their body positions during the collaboration process. **(3) Only the listener's face was tracked.** This error may happen due to the same reason as condition 2. **(4) Neither of the learners' faces was tracked.** This error may happen due to both learners being out of the camera (e.g., disengaged and talking to other classmates) or the presence of occlusion.

We used the OpenFace 2.0 facial behavior analysis toolkit [36] to count how many times each of the four conditions occur in our dataset. OpenFace supports automatic facial recognition by generating the number of detected *face_ids* as well as the location of the head in the horizontal axis *pose_Tx* for every video frame. For condition 1, the number of *face_ids* is 2; For conditions 2 and condition 3, the number of *face_ids* is 1; for condition 4, the number of *face_ids* is 0. We then used the *pose_Tx* to differentiate condition 2 and condition 3. Table 3 shows the number of video segments in the dataset that fall into each face-tracking condition.

Table 3: Number of video segments of each tracking condition

Condition	Confusion	Conflict	Other
1: both faces tracked	271	473	4,616
2: only speaker's face tracked	77	183	1,510
3: only listener's face tracked	68	169	1,643
4: both faces mistracked	52	99	783
Total	467	924	8,552

5 FEATURE EXTRACTION

5.1 Language Features

To extract linguistic features, we represented each spoken utterance with TF∗IDF, BERT, and RoBERTa. Given that signal words or phrases (e.g., "*confuse*", "*do not*", "*not know*") appear frequently when learners express their confusion or conflict, we generated Tf∗IDF embeddings. We also used BERT and RoBERTa to generate the semantic embedding of each utterance. RoBERTa is a variant of BERT trained on longer sequences. RoBERTa dynamically changes the masking pattern applied to the training data and has outperformed BERT on a series of language processing tasks, such as machine translation and question answering [24]. We used the Hugging Face bert-base-uncased [3] and xlm-roberta-based [42] models to generate a 768-dimensional language embedding for each utterance.

5.2 Audio Features

We used openSMILE, an open-source automatic acoustic feature extraction toolkit [12] for extracting acoustic-prosodic indicators with a 20ms frame and a window shift of 10ms. For each 20 ms of audio, 25 eGeMAPS [11] low-level descriptors (LLDs) were generated, including loudness (1 feature), pitch (10 features), and mel frequency cepstral coefficients (MFCCs) (4 features). Following an established downsampling strategy [38], we averaged LLDs every 2 consecutive frames. Apart from acoustic-prosodic LLDs, we also experimented with Wav2Vec, a transformer-based audio embedding network that has shown state-of-the-art performance on a series of speech-related tasks, such as speech recognition [55] and speech emotion recognition [38]. We used a Wav2Vec-based [57] model to generate a 768-dimensional audio embedding for each audio segment.

Table 4: Results for selecting best-performing unimodal features. Label distribution: Confusion (4.7%), Conflict (9.3%), and Other (86.0%). P: Precision, R: Recall, F: F-1 Score, A: Overall Accuracy.

		Confusion			Conflict			Other			
Modality	**Unimodal Features**	**P**	**R**	**F**	**P**	**R**	**F**	**P**	**R**	**F**	**A**
Language	TF*IDF	0.34	0.48	0.40	0.25	0.37	0.29	0.87	0.80	0.83	0.77
	BERT	0.46	0.61	0.52	0.51	0.58	0.55	0.92	0.91	0.91	0.87
	RoBERTA	0.53	0.64	**0.57**	0.55	0.65	**0.61**	0.93	0.88	0.91	**0.89**
Audio	Loudness	0.13	0.10	0.11	0.17	0.26	0.20	0.86	0.83	0.85	0.62
	Pitch	0.18	0.10	0.13	0.10	0.17	0.14	0.81	0.93	0.86	0.63
	MFCCs	0.18	0.34	0.25	0.26	0.37	0.28	0.90	0.70	0.80	0.66
	Wav2Vec	0.25	0.45	**0.32**	0.27	0.49	**0.36**	0.87	0.88	0.88	**0.70**
Video	Eye Gaze	0.17	0.32	0.23	0.33	0.24	0.29	0.85	0.82	0.83	0.68
	Head Pose	0.19	0.27	0.22	0.31	0.21	0.27	0.88	0.85	0.86	0.65
	Facial AUs	0.40	0.54	**0.46**	0.31	0.39	**0.35**	0.89	0.86	0.88	**0.75**

5.3 Video Features

In this paper, we used OpenFace, 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 [29]. 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.

6 EXPERIMENTS AND RESULTS

Given that the feature space (relative to the dataset used in this study) is large, we first reduced the feature space by identifying and removing weakly relevant or irrelevant features. To do this, we provided every feature extracted from language, video, and video modalities as the input to a multilayer-perceptron (MLP) classifier with an embedding size of 128 for two linear layers; two dropout layers with a rate of 0.5 were added to each linear layer to alleviate over-fitting. The *Softmax* activation function was used in the last output layer to output the classification results. We used the synthetic minority oversampling technique (SMOTE) [8] to mitigate the negative influence of the imbalanced label distribution of confusion (4.7%) and conflict (9.3%) within our dataset. SMOTE was only performed on the training set, and class distributions for the validation and testing sets were left unchanged.

We conducted five-fold cross-validation to train and validate the models. We used an Adam optimizer [21] with the learning rate of $1 \times e^{-3}$ to train the classification model up to 100 epochs. We evaluated the trained model using an F-1 score [13] combined from precision and recall for each one of the three classes. Table 4 shows the confusion and conflict classification performance trained on each of the single features. From the results in the table, we identified the best-performing unimodal features in each modality: fine-tuned RoBERTa, Wav2Vec, and facial AUs in the language, audio, and video modality, respectively.

6.1 Investigating the Impact of Unimodal Data Noise

We first consider RQ1: *To what extent do language errors, audio noise, and facial mistracking impact the accuracy of modeling confusion and conflict during collaborative learning?* For different data noise levels within each modality, we built and compared the performance of several supervised unimodal models trained on the best-performing feature identified in Table 4; we used the same experimental setup and training strategies as described above.

6.1.1 ***Impact of Word Error Rate.*** Figure 2 shows the performance of unimodal models using language-derived features under different WER levels. In the figure, as WER increased from 0 to 0.2, the overall accuracy, as well as the F-1 scores for both confusion and conflict, remained relatively stable. When WER increased from 0.2 to 0.3, the performance suffered a drastic degradation, with the overall accuracy decreasing from 0.84 to 0.73, and the F-1 scores decreasing from 0.55 to 0.40 for confusion and from 0.53 to 0.37 for conflict.

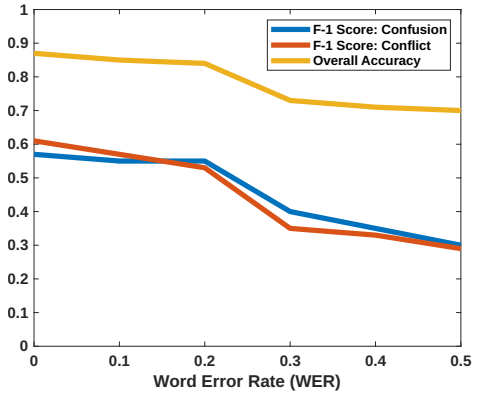

Figure 2: Unimodal models using language-derived features

6.1.2 ***Impact of Audio Noise.*** Figure 3 shows the performance of unimodal models both trained and tested using audio-derived features under different SNR levels. As shown in the figure, the performance was very sensitive to noise, with performance decreasing

sharply as the noise level increased. From $5 < SNR$ to $0 < SNR < 5$, overall accuracy decreased from 0.79 to 0.61; the F-1 scores for confusion decreased from 0.39 to 0.19, and for conflict decreased from 0.48 to 0.36. The continued to decrease drastically as SNR declined.

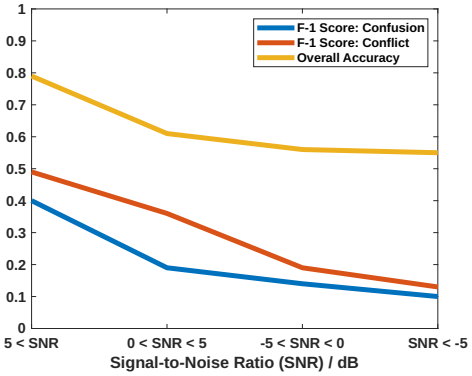

Figure 3: Unimodal models using audio-derived features

6.1.3 ***Impact of Facial Recognition Errors.*** Figure 4 shows the performance of unimodal models both trained and tested using video-derived features under different facial tracking conditions. We considered four conditions: (1) both learners' faces were tracked; (2) only the speaker's face was tracked; (3) only the listener's face was tracked, and (4) neither of the learners' faces was tracked. As shown in the figure, the model yielded the highest accuracy of 0.79 when both learners' faces were successfully tracked (condition 1). Surprisingly, the model performed comparably even when only the listener's face was tracked (condition 3), with a degraded accuracy of 0.71, and F-1 scores declined from 0.51 to 0.44 for confusion and 0.41 to 0.33 for conflict between condition 1 and condition 3. We did not investigate condition 4 because there were no facial features generated by OpenFace when both learners' faces were missing.

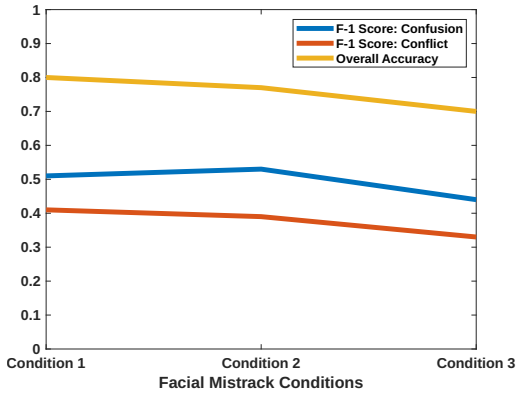

Figure 4: Unimodal models using video-derived features

6.2 Examining the Performance of Multimodal Modeling

We next consider RQ2: *To what extent does integrating language, audio, and facial behaviors compensate for the negative impact of unimodal data noise?* To answer this question, we built several supervised multimodal models trained on the fusion of best-performing features in the language, audio, and video modalities. Figure 5 shows the overview of the multimodal model architecture. The other multimodal models followed the same structure with a subset of the language, audio, and video modalities.

In this study, we experimented with two types of feature fusion methods [2]: model-agnostic and model-based. For model-agnostic methods, we used early and late fusion. Early fusion concatenated unimodal features after applying z-score normalization. Late fusion trained separate models with separate unimodal features and calculated the numerical average of their outputs. For model-based methods, we experimented with two neural-network-based fusion approaches: tensor fusion [59] and cross-attention fusion [31]. Tensor fusion transforms multimodal features into a 3D feature tensor, while cross-attention fusion uses a shared transformer encoder to attend to different modalities.

We then set noise thresholds for selecting subsets of the data where the accuracy of each unimodal model decreased drastically in association with noise. We selected audio segments with a signal-to-noise ratio (SNR) of less than +5 dB, as the model's accuracy drastically decreased beyond this point. Similarly, we used the video segments that did not track either learner's face, as model accuracy was the lowest in this face-tracking condition. Then, we introduced extra modalities to this baseline unimodal model to examine if multimodal models outperform the unimodal baselines trained with noisy data from each single modality.

First, we tested the impact of integrating multimodal data on compensating for *noisy language* using a baseline model with a WER of 0.3. The baseline model's accuracy was 0.73, with F-1 scores of 0.40 for confusion and 0.37 for conflict. We then constructed several multimodal models (as described above) with additive audio and video data. All except the late fusion model performed better than the baseline, with higher overall accuracy. Cross-attention fusion achieved the best performance by integrating language, audio, and video data together, with the highest accuracy of 0.80, an F-1 score of 0.46 for confusion, and an F-1 score of 0.48 for conflict. Second, we tested the impact of integrating multimodal data toward compensating for *noisy audio.* A baseline model with a SNR < +5 dB achieved an accuracy of 0.61, with F-1 scores of 0.19 for confusion and 0.38 for conflict. We then constructed several multimodal models with additive language and video data, where all multimodal models performed better than the baseline, with higher overall accuracy. Finally, we tested the impact of integrating multimodal data on compensating for *incomplete video* by experimenting with a baseline model using video segments that mistracked at least one learner. The baseline model's accuracy was 0.73, with F-1 scores of 0.46 for confusion and 0.32 for conflict. We then constructed several multimodal models with additive language and video data, where all multimodal models performed better than the baseline, with higher overall accuracy. Table 5 shows the performance of multimodal models when single modalities involve data noise.

7 DISCUSSION

This study has investigated the impact of data noise on multimodal recognition of confusion and conflict moments during dyads of

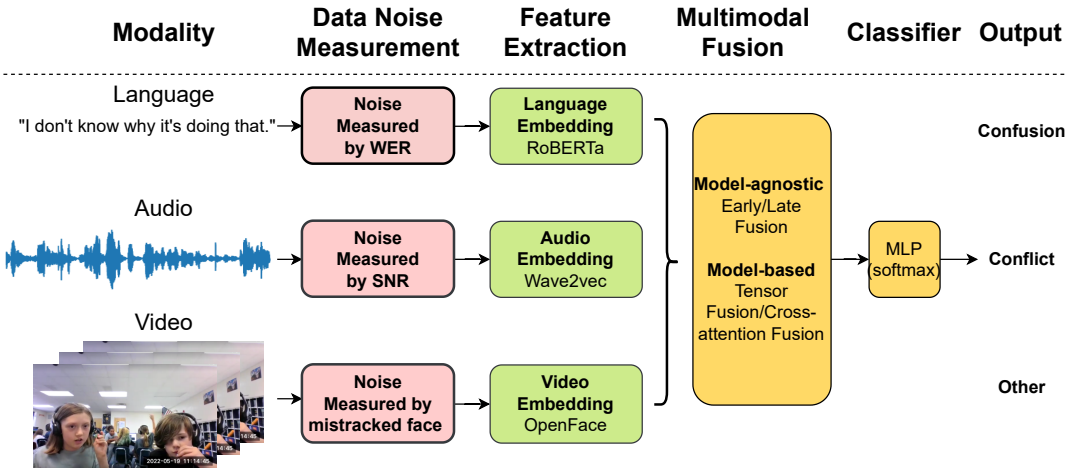

Figure 5: Architecture of multimodal modeling using language, audio, and video data.

Table 5: F-1 scores and accuracy for multimodal models.

	Late Fusion			Early Fusion			Tensor Fusion			Cross-Attention Fusion		
	Confusion	Conflict	Acc	Confusion	Conflict	Acc	Confusion	Conflict	Acc	Confusion	Conflict	Acc
Baseline: Noisy Language (WER = 0.3)	Confusion 0.40, Conflict 0.37, Acc 0.73											
Noisy Language + Audio	0.33	0.38	0.71	0.40	0.41	0.75	0.40	0.43	0.74	0.42	0.46	0.75
Noisy Language + Video	0.35	0.36	0.72	0.43	0.37	0.74	0.44	0.40	0.75	0.45	0.40	0.75
Noisy Language + Audio + Video	0.38	0.41	0.74	0.42	0.43	0.78	0.43	0.45	0.78	**0.46**	**0.48**	**0.80**
Baseline: Noisy Audio (SNR <5 dB)	Confusion 0.19, Conflict 0.38, Accuracy 0.61											
Noisy Audio + Language	0.51	0.45	0.65	0.55	0.53	0.80	0.58	0.55	0.80	0.60	0.55	0.81
Noisy Audio + Video	0.45	0.35	0.63	0.48	0.38	0.71	0.48	0.40	0.72	0.50	0.41	0.74
Noisy Audio + Language + Video	0.55	0.48	0.78	0.61	0.55	0.84	0.61	0.57	0.85	**0.65**	**0.60**	**0.88**
Baseline: Mistracked Videos (Condition 2, 3, 4)	Confusion 0.46, Conflict 0.32, Accuracy 0.73											
Mistracked Video + Language	0.53	0.45	0.81	0.58	0.60	0.88	0.58	0.62	0.87	0.61	0.65	0.91
Mistracked Video + Audio	0.45	0.40	0.74	0.45	0.40	0.80	0.45	0.41	0.80	0.47	0.40	0.84
Mistracked Video + Language + Audio	0.48	0.51	0.78	0.60	0.62	0.86	0.64	0.63	0.88	**0.65**	**0.68**	**0.92**

learners' collaborative learning activities. This section discusses the experimental results with respect to two research questions.

7.1 The Impact of Unimodal Data Noise

7.1.1 ***Noisy Language.*** The experimental results showed that unimodal models using language-derived features trained on noisy transcripts up to a *WER* of 0.2 could perform comparably well as models trained with clean transcripts manually generated by humans. Transcripts with a *WER* of 0.3 were too noisy, and the model performance suffered a drastic degradation. In another study, Southwell et al. [47] conducted extensive experiments to compare three widely adopted ASR engines (Google, Rev.ai [40], and IBM Watson) on transcribing audio recordings of middle-school students engaged in small group work. The authors found that it was extremely difficult to obtain serviceable transcripts by current (2023) ASR engines, which generated overall high *WERs* on this task (0.84 - 0.95). These findings highlight the main challenge of deploying intelligent systems to support collaboration in real-world classroom environments: obtaining serviceable transcriptions of student discourse. Indeed, on our corpus, cloud-based ASR performed similarly poorly, with Google Speech-to-text [52] having an average *WER* of 0.78 (SD = 0.54), and those generated by IBM Watson [56] have an average *WER* of 0.89 (SD = 0.46). This finding led to our decision to perturb manual transcripts for the purposes of experimentation.

7.1.2 ***Noisy Audio.*** Overall, the audio recordings in our dataset are noisy, with an average *SNR* of +1.3 dB. The experimental results showed that the performance of unimodal models using audio-derived features was very sensitive to noise and showed a steep degradation as soon as the audio *SNR* decreased below +5 dB, where the overall accuracy decreased drastically from 0.79 to 0.61, the F-1 score of confusion from 0.40 to 0.19, and the F-1 score of conflict from 0.48 to 0.39. These results suggest that audio data with *SNR* of +5 dB could potentially be considered acceptable. It is challenging to collect audio of this quality in real classroom environments, where *SNR* usually ranges from -7 dB to +5 dB [18]. Quality can be improved when learners use headsets and wear noise-canceling microphones close to their mouths, but a tradeoff is that the headsets detract from the fluid interplay of individual, small group, and whole class discourse.

7.1.3 ***Incomplete Video***. The experimental results showed that unimodal models using video-derived features trained with video segments tracking at least one learner's face in the pair could still perform comparably well to unimodal models trained with video segments tracking both learners' faces (an accuracy of 0.71 versus 0.79). It is not surprising that when a learner expresses confusion or conflict, it can be detected through the speaker's face. However, our results found that reasonable classification accuracy can also be achieved even when only one learner's face is tracked. This finding is consistent with a recent study by Järvenoja et al. [19] where the authors investigated how socially shared emotions emerged during collaborative learning activities. Taken together, it appears that tracking at least a sub-group of learners' faces is sufficient for recognizing a speaker's confusion and or conflict, but a high-resolution wide-angle camera is still recommended.

7.2 The Effect of Multimodal Modeling

This study trained several multimodal models and found that when there is data noise present in a single modality, fusing information from other modalities can effectively compensate for the negative impact of unimodal data noise, ultimately leading to better accuracy in recognizing a speaker's confusion and conflict. Specifically for *noisy language*, introducing additional audio and facial information could enhance model accuracy from 0.73 to 0.80. This improvement in accuracy indicates that audio and video data can also offer valuable insights into detecting confusion and conflict. The addition of audio and video data can reveal non-verbal cues, such as tone and facial expressions, which are often absent from text-based data. This information can provide additional context, helping the model to better understand the situation and make more accurate predictions. Similarly, for *noisy audio*, fusing additional language and facial information could enhance model accuracy from 0.61 to 0.88; for *incomplete video*, fusing additional language and audio information could improve model accuracy from 0.73 to 0.92.

However, in a real data collection environment, a higher level of noise in the audio data will negatively impact both speech-to-text translation and prosody analysis. For speech-to-text translation, a higher level of audio noise can make it more difficult for ASR engines to accurately transcribe the speech. This can lead to a higher *WER* in the resulting transcript [20]. Similarly, a higher level of audio noise can make it more difficult to accurately identify the audio features, such as patterns of pitch [28]. Hence, the setup of intelligent systems in classrooms should aim to capture clean speech, if possible, either by using advanced microphones or separating learner groups to avoid ambient sound. To improve performance in transcribing noise speech, recent pre-trained ASR models can be fine-tuned.

In our comparison of multimodal models trained with various fusion techniques in the presence of noise, the results showed that neural-network-based fusion approaches generally achieved higher confusion and conflict recognition accuracy than traditional early and late fusion approaches. The main advantage of neural-network-based fusion approaches over model-agnostic-based fusion approaches lies in their ability to exploit the underlying relationships and mutual information among modalities [2, 22]. Hence, the experimental results suggest that in the face of the same data noise level, model-based fusion approaches could have more robust performance in modeling collaborative learning than model-agnostic fusion approaches.

7.3 Limitations and Future Work

The current study has important limitations. First, our random perturbation approach to generate text errors may not fully simulate ASR errors, as audio transcribed with an ASR engine can have noisy text that is phonetically similar to the reference text. Second, the dataset was relatively small, consisting of recordings from just 25 learner dyads, so the acceptable noise level identified here may not be generalizable to learners in other age groups or learning environments, such as online learning. In addition, the noise levels experimented with in this study were not fine-grained enough. We generated noisy language data with a *WER* granularity of 0.1; we split noisy audio data with a *SNR* granularity of 5 dB. Using smaller granularities in future studies will provide more practical implications for deploying intelligent systems in actual classroom environments. Last, this study stopped short of attempting to improve the performance of the state-of-the-art multimodel fusion models tested. Toward this goal, we are currently developing an intelligent system that adopts an adaptive fusion strategy, in which information from different modalities is dynamically integrated based on the estimation of their noise level so that more informative modalities are prioritized to improve multimodal modeling performance over time.

8 CONCLUSION

Intelligent systems hold great promise to support collaborative learning, but the noise that accompanies the data poses great challenges to analyzing and understanding learners' dialogues. It is important to develop effective methods for intelligent systems to perform robustly in the face of substantial noise. This paper takes a first step toward addressing this challenge by understanding the impact of noisy language, noisy audio, and incomplete video on modeling learners' interactions during collaborative activities.

The results of extensive experiments showed that in the language modality, the model's accuracy for detecting confusion and conflict decreased drastically when the *WER* exceeded 20%. In the audio modality, the model's accuracy decreased sharply when the *SNR* dropped below 5 dB. In the video modality, the model's accuracy remained relatively constant at a comparable level as long as at least one learner's face was successfully tracked. To further investigate the effect of integrating multimodal data, given the presence of unimodal data noise, we trained several multimodal models. The results showed that combining other modalities of data could effectively compensate for the negative effect of noise in unimodal data, ultimately leading to improved modeling accuracy.

ACKNOWLEDGMENTS

This research was supported by the National Science Foundation through grants DRL-2229612 and DRL-1721160. 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.

REFERENCES

[1] Oluremi B Ayoko, Victor J Callan, and Charmine EJ Härtel. 2008. The influence of team emotional intelligence climate on conflict and team members' reactions to conflict. *Small Group Research* 39, 2 (2008), 121–149.

[2] Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. 2018. Multimodal machine learning: A survey and taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 41, 2 (2018), 423–443.

[3] Bert-based. 2023. https://huggingface.co/bert-base-uncased.

[4] Natalia Bogach, Elena Boitsova, Sergey Chernonog, Anton Lamtev, Maria Lesnichaya, Iurii Lezhenin, Andrey Novopashenny, Roman Svechnikov, Daria Tsikach, Konstantin Vasiliev, et al. 2021. Speech processing for language learning: A practical approach to computer-assisted pronunciation teaching. *Electronics* 10, 3 (2021), 235.

[5] Dan Bohus and Eric Horvitz. 2010. Facilitating multiparty dialog with gaze, gesture, and speech. In *Proceedings of the International Conference on Multimodal Interfaces and the Workshop on Machine Learning for Multimodal Interaction*. 1–8.

[6] Hervé Bourlard, Hynek Hermansky, and Nelson Morgan. 1996. Towards increasing speech recognition error rates. *Speech Communication* 18, 3 (1996), 205–231.

[7] Mehmet Celepkolu and Kristy Elizabeth Boyer. 2018. The importance of producing shared code through pair programming. In *Proceedings of the 49th ACM Technical Symposium on Computer Science Education (SIGCSE 2018)*. 765–770.

[8] Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. 2002. SMOTE: synthetic minority over-sampling technique. *Journal of artificial intelligence research* 16 (2002), 321–357.

[9] Yong Dich, Joseph Reilly, and Bertrand Schneider. 2018. Using physiological synchrony as an indicator of collaboration quality, task performance and learning. In *Proceedings of the 19th International Conference on Artificial Intelligence in Education (AIED 2018)*. Springer, 98–110.

[10] Lucca Eloy, Angela EB Stewart, Mary Jean Amon, Caroline Reinhardt, Amanda Michaels, Chen Sun, Valerie Shute, Nicholas D Duran, and Sidney D'Mello. 2019. Modeling team-level multimodal dynamics during multiparty collaboration. In *Proceedings of the 21st International Conference on Multimodal Interaction (ICMI 2019)*. 244–258.

[11] Florian Eyben, Klaus R Scherer, Björn W Schuller, Johan Sundberg, Elisabeth André, Carlos Busso, Laurence Y Devillers, Julien Epps, Petri Laukka, Shrikanth S Narayanan, et al. 2015. The Geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing. *IEEE Transactions on Affective Computing* 7, 2 (2015), 190–202.

[12] Florian Eyben, Martin Wöllmer, and Björn Schuller. 2010. Opensmile: the munich versatile and fast open-source audio feature extractor. In *Proceedings of the 18th International Conference on Multimedia (ICME 2010)*. 1459–1462.

[13] Cyril Goutte and Eric Gaussier. 2005. A probabilistic interpretation of precision, recall and F-score, with implication for evaluation. In *Proceedings of the 27th European Conference on IR Research (ECIR 2005)*. Springer, 345–359.

[14] Reihaneh H Hariri, Erik M Fredericks, and Kate M Bowers. 2019. Uncertainty in big data analytics: survey, opportunities, and challenges. *Journal of Big Data* 6, 1 (2019), 1–16.

[15] Nathan Henderson, Wookhee Min, Jonathan Rowe, and James Lester. 2020. Enhancing affect detection in game-based learning environments with multimodal conditional generative modeling. In *Proceedings of the 22nd International Conference on Multimodal Interaction (ICMI 2020)*. 134–143.

[16] Jay Heo, Hae Beom Lee, Saehoon Kim, Juho Lee, Kwang Joon Kim, Eunho Yang, and Sung Ju Hwang. 2018. Uncertainty-aware attention for reliable interpretation and prediction. *Advances in Neural Information Processing Systems* 31 (2018).

[17] Markus Höferlin, Benjamin Höferlin, Daniel Weiskopf, and Gunther Heidemann. 2011. Uncertainty-aware video visual analytics of tracked moving objects. *Journal of Spatial Information Science* 2011, 2 (2011), 87–117.

[18] Clare S Howard, Kevin J Munro, and Christopher J Plack. 2010. Listening effort at signal-to-noise ratios that are typical of the school classroom. *International journal of audiology* 49, 12 (2010), 928–932.

[19] Hanna Järvenoja, Sanna Järvelä, and Jonna Malmberg. 2020. Supporting groups' emotion and motivation regulation during collaborative learning. *Learning and Instruction* 70 (2020), 101090.

[20] James Kennedy, Séverin Lemaignan, Caroline Montassier, Pauline Lavalade, Bahar Irfan, Fotios Papadopoulos, Emmanuel Senft, and Tony Belpaeme. 2017. Child speech recognition in human-robot interaction: evaluations and recommendations. In *Proceedings of the 2017 International Conference on Human-robot Interaction*. 82–90.

[21] Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).

[22] Dana Lahat, Tülay Adali, and Christian Jutten. 2015. Multimodal data fusion: an overview of methods, challenges, and prospects. *Proc. IEEE* 103, 9 (2015), 1449–1477.

[23] Emily R Lai. 2011. Collaboration: A literature review. *Pearson Publisher. Retrieved November* 11 (2011), 2016.

[24] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692* (2019).

[25] Yingbo Ma, Mehmet Celepkolu, and Kristy Elizabeth Boyer. 2022. Detecting Impasse During Collaborative Problem Solving with Multimodal Learning Analytics. In *Proceedings of the 12th International Learning Analytics and Knowledge Conference (LAK 2022)*. 45–55.

[26] Yingbo Ma, Joseph B Wiggins, Mehmet Celepkolu, Kristy Elizabeth Boyer, Collin Lynch, and Eric Wiebe. 2021. The Challenge of Noisy Classrooms: Speaker Detection During Elementary Students' Collaborative Dialogue. In *Proceedings of the 22nd International Conference on Artificial Intelligence in Education (AIED 2021)*. Springer, 268–281.

[27] Ioannis Magnisalis, Stavros Demetriadis, and Anastasios Karakostas. 2011. Adaptive and intelligent systems for collaborative learning support: A review of the field. *IEEE transactions on Learning Technologies* 4, 1 (2011), 5–20.

[28] Kasiprasad Mannepalli, Panyam Narahari Sastry, and Maloji Suman. 2018. Analysis of emotion recognition system for Telugu using prosodic and formant features. In *Proceedings of the Speech and Language Processing for Human-Machine Communications (CSI 2015)*. Springer, 137–144.

[29] Brais Martinez, Michel F Valstar, Bihan Jiang, and Maja Pantic. 2017. Automatic analysis of facial actions: A survey. *IEEE transactions on affective computing* 10, 3 (2017), 325–347.

[30] Robert G Moulder, Nicholas D Duran, and Sidney K D'Mello. 2022. Assessing Multimodal Dynamics in Multi-Party Collaborative Interactions with Multi-Level Vector Autoregression. In *Proceedings of the 24th International Conference on Multimodal Interaction (ICMI 2022)*. 615–625.

[31] Arsha Nagrani, Shan Yang, Anurag Arnab, Aren Jansen, Cordelia Schmid, and Chen Sun. 2021. Attention bottlenecks for multimodal fusion. *Advances in Neural Information Processing Systems* 34 (2021), 14200–14213.

[32] Nicholas J Napoli, Chad L Stephens, Kellie D Kennedy, Laura E Barnes, Ezequiel Juarez Garcia, and Angela R Harrivel. 2023. NAPS Fusion: A framework to overcome experimental data limitations to predict human performance and cognitive task outcomes. *Information Fusion* 91 (2023), 15–30.

[33] Jauwairia Nasir, Aditi Kothiyal, Barbara Bruno, and Pierre Dillenbourg. 2021. Many are the ways to learn identifying multi-modal behavioral profiles of collaborative learning in constructivist activities. *International Journal of Computer-Supported Collaborative Learning* 16, 4 (2021), 485–523.

[34] Piia Näykki, Sanna Järvelä, Paul A Kirschner, and Hanna Järvenoja. 2014. Socio-emotional conflict in collaborative learning—A process-oriented case study in a higher education context. *International Journal of Educational Research* 68 (2014), 1–14.

[35] Jennifer K Olsen, Kshitij Sharma, Nikol Rummel, and Vincent Aleven. 2020. Temporal analysis of multimodal data to predict collaborative learning outcomes. *British Journal of Educational Technology* 51, 5 (2020), 1527–1547.

[36] OpenFace. 2023. https://github.com/TadasBaltrusaitis/OpenFace.

[37] George Papandreou, Athanassios Katsamanis, Vassilis Pitsikalis, and Petros Maragos. 2009. Adaptive multimodal fusion by uncertainty compensation with application to audiovisual speech recognition. *IEEE Transactions on Audio, Speech, and Language Processing* 17, 3 (2009), 423–435.

[38] Leonardo Pepino, Pablo Riera, and Luciana Ferrer. 2021. Emotion recognition from speech using wav2vec 2.0 embeddings. *arXiv preprint arXiv:2104.03502* (2021).

[39] Sarah E Peterson and Jeffrey A Miller. 2004. Comparing the quality of students' experiences during cooperative learning and large-group instruction. *The Journal of Educational Research* 97, 3 (2004), 123–134.

[40] Rev.ai. 2023. https://docs.rev.ai/resources/tutorials/get-started-python/.

[41] Josue Reyes-Garcia, Hiram Galeana-Zapién, Alejandro Galaviz-Mosqueda, and Cesar Torres-Huitzil. 2018. Evaluation of the impact of data uncertainty on the prediction of physiological patient deterioration. *IEEE Access* 6 (2018), 38595–38606.

[42] RoBERTa. 2023. https://huggingface.co/xlm-roberta-base.

[43] Fernando J Rodríguez, Kimberly Michelle Price, and Kristy Elizabeth Boyer. 2017. Expressing and addressing Uncertainty: A study of collaborative Problem-solving dialogues. Philadelphia, PA: International Society of the Learning Sciences.

[44] Sanja Seljan and Ivan Dunđer. 2014. Combined automatic speech recognition and machine translation in business correspondence domain for english-croatian. *International Journal of Industrial and Systems Engineering* 8, 11 (2014), 1980–1986.

[45] Silero. 2023. https://github.com/snakers4/silero-models.

[46] Snap! 2023. https://snap.berkeley.edu/.

[47] Rosy Southwell, Samuel Pugh, E Margaret Perkoff, Charis Clevenger, Jeffrey B Bush, Rachel Lieber, Wayne Ward, Peter Foltz, and Sidney D'Mello. 2022. Challenges and Feasibility of Automatic Speech Recognition for Modeling Student Collaborative Discourse in Classrooms. In *Proceedings of The 15th International Conference on Educational Data Mining (EDM 2022)*. ERIC.

[48] 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.

[49] Shree Krishna Subburaj, Angela EB Stewart, Arjun Ramesh Rao, and Sidney K D'Mello. 2020. Multimodal, multiparty modeling of collaborative problem solving performance. In *Proceedings of the International Conference on Multimodal Interaction (ICMI 2020)*. 423–432.

[50] Zheng-Hua Tan and Børge Lindberg. 2010. Low-complexity variable frame rate analysis for speech recognition and voice activity detection. *IEEE Journal of Selected Topics in Signal Processing* 4, 5 (2010), 798–807.

[51] Chinchu Thomas and Dinesh Babu Jayagopi. 2017. Predicting student engagement in classrooms using facial behavioral cues. In *Proceedings of the 1st ACM SIGCHI International Workshop on Multimodal Interaction for Education*. 33–40.

[52] Google Speech to Text. 2023. https://cloud.google.com/speech-to-text.

[53] Jennifer Tsan, Jessica Vandenberg, Zarifa Zakaria, Danielle C Boulden, Collin Lynch, Eric Wiebe, and Kristy Elizabeth Boyer. 2021. Collaborative dialogue and types of conflict: An analysis of pair programming interactions between upper elementary students. In *Proceedings of the 52nd ACM Technical Symposium on Computer Science Education (SIGCSE 2021)*. 1184–1190.

[54] 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 10th International Conference on Learning Analytics & Knowledge (LAK 2020)*. 295–304.

[55] Yiming Wang, Jinyu Li, Heming Wang, Yao Qian, Chengyi Wang, and Yu Wu. 2022. Wav2vec-switch: Contrastive learning from original-noisy speech pairs for robust speech recognition. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 7097–7101.

[56] Watson. 2023. https://www.ibm.com/watson.

[57] Wave2vec. 2023. https://github.com/facebookresearch/fairseq/tree/main/examples/wav2vec.

[58] OpenAI Whisper. 2023. https://github.com/openai/whisper.

[59] Amir Zadeh, Minghai Chen, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2017. Tensor fusion network for multimodal sentiment analysis. *arXiv preprint arXiv:1707.07250* (2017).

[60] Zarifa Zakaria, Jessica Vandenberg, Jennifer Tsan, Danielle Cadieux Boulden, Collin F Lynch, Kristy Elizabeth Boyer, and Eric N Wiebe. 2022. Two-computer Pair Programming: Exploring a Feedback Intervention to Improve Collaborative Talk in Elementary Students. *Computer Science Education* 32, 1 (2022), 3–29.

[61] Patryk Żywica. 2020. Application of uncertainty-aware similarity measure to classification in medical diagnosis. In *Proceedings of the 2020 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*. IEEE, 1–8.