Inferring the climate in classrooms from audio and video recordings: A machine learning approach

Anusha James
Nanyang Technological University
anush.james@gmail.com

Mohan Kashyap
Nanyang Technological University
mkpargi@ntu.edu.sg

Victoria Chua
Nanyang Technological University
vicchuayh@gmail.com

Tomasz Maszczyk
Nanyang Technological University
tomasz.maszczyk@gmail.com

Ana Moreno Núñez
National Institute of Education
anamorenonunezphd@gmail.com

Rebecca Bull
National Institute of Education
rebecca.bull@nie.edu.sg

Justin Dauwels
Nanyang Technological University
jdauwels@ntu.edu.sg

***Abstract*— The classroom climate is shaped by a combination of teacher practices and peer relationships. The Classroom Assessment Scoring System (CLASS) has been designed to observe and code classroom interactions between students and teachers in order to provide formative feedback on teaching practices and improve teacher instruction. But the turnover time for training, observing and coding makes it hard to generate instant feedback. Since there are few automated assessment tools designed to measure the classroom climate, we propose a novel system for automatic assessment of classroom climate, based on speech, behavioural cues and video features by applying machine learning techniques. This paper elaborates on the design and validation of an audio-video analytics platform for predicting classroom climate. Employing machine learning classifiers instead of subjective measures can ease and expedite the coding. We presume our system can empower education systems to continuously review and improve teaching strategies thus promoting smart classroom in the future.**

Keywords—classroom climate prediction, machine learning, audio-video analytics, social behavior, educational research

I. INTRODUCTION

There is growing interest in classroom climate as it has been found to be related to positive student outcomes such as academic achievement, emotional development and social competence. The classroom climate, assessed by CLASS(Classroom Assessment Scoring System), can be positive or negative. The CLASS [1] is a system for gauging interactional quality between students and teachers in preschool classrooms. The measure consists of 10-subscales, of which positive climate and negative climate are the two subscales. Classroom climate refers to the prevailing mood, attitudes and emotional tone of the classroom. A positive climate classroom feels secure, hospitable, respectful and supportive of constructive learning process whereas negative climate reflects negativity such as hostility, chaos, or aggression. Classroom climate does not just happen - it is created, and an effective teacher can deliberately shape a classroom to a positive learning environment. There is a need to focus on teaching styles which emphasizes fostering positive and supportive classroom interactions among teacher and students. Rapport established between two parties can be detected through speech patterns and body movements, thus kindling our idea of leveraging speech analysis and non-verbal behaviour including facial expressions to investigate classroom interactions.

Existing observational measures to evaluate teachers such as CLASS are coded manually by trained professionals after observing a multitude of classroom sessions. This yields day-to-day qualitative and formative feedback to fine-tune their teaching strategies. However, this process, from training to observation and assessment, is tedious, labour-intensive and expensive. Therefore, finding objective indicators that correlate with CLASS dimensions, and can be easily and quickly extracted, would be helpful for teacher professional development and decreasing turnover time.

In this paper, we explore the automated analysis of teaching practices by applying speech processing and video processing technologies. We showcase the feasibility of applying artificial intelligence to help create positive learning environments through assisting in the evaluation process of teaching quality, making decisions regarding effective teaching strategies, and enhancing students' learning outcomes. Our study is based on audio-visual recordings of preschool classrooms. The designed pipeline extracts speech and video features of teachers and students, followed by the application of machine learning techniques to infer the classroom climate automatically. We were able to distinguish between positive and negative CLASS climate scores with 70-78% weighted F_1-score metrics (estimated by 10-fold cross-validation).

The rest of the paper is organized as follows. In Section II, we briefly summarize the relevant work, while in Section III, we discuss about the data and challenges that actuated this pipeline. In Section IV, we provide information on our proposed approach and algorithms. In Section V, we present our numerical results, followed by Section VI, in which we describe

directions of future work.

II. Related Works

To date, the majority of automated assessment studies has concentrated on creating tools that assess students or teachers when they are interacting with Computer-Based Educational Systems (CBESs), of which the entirety of interaction happens in front of a computer or tablet. However, there has been increasing interest among researchers and practitioners in capturing, modelling and analysing, learning and teaching experiences beyond computer-based learning environments [2-4]. These studies focus on learning processes that take place face-to-face, in co-located spaces, such as classrooms that utilise project-based learning, collaborative problem solving, embodied interaction, body-based learning, or simply, traditional teacher-students classrooms.

For instance, some studies focus on leveraging multimodal data streams of student speech and hand gestures from video to investigate learning processes and reasoning techniques [2] and predict students' expertise [3] and team success [5] within small groups of students working in hands-on project-based learning environments. Similar approaches have been applied to study students' learning processes and social dynamics in collaborative mathematics tasks. In a study in [6], they utilised audio transcripts and video recordings of students' posture, calculator usage and movement to predict success at solving mathematics problems and expertise levels, while [7] found that prosodic features and hand gestures could differentiate level of leadership and expertise of high school students collaborating on mathematics problems. Understandably, researchers have also explored the application of video and audio analytics in classrooms that utilise body-based learning and embodied interaction to derive meaningful data about students' learning processes and social dynamics. Andrade and colleagues [8] were able to predict the type of reasoning strategy adopted by students (aged 6 to 7 years old) based on body posture, eye gaze, hand gestures and audio recordings. Audio transcripts of students' speech and students' hand gestures were employed to identify learning patterns in a body-based learning task within fourth graders [9]. In [21] video analytics and machine learning techniques were applied to detect students' affect from facial expressions and gross body movements during interactions with an educational physics game. Albeit the focus on small groups of students, instead on a classroom level, these studies indicate the power of harnessing audio and video analytics to aid student assessment and understand learning processes in the classroom. In our study, we investigate whether similar approaches of harnessing video and audio analytics could be applied to the classroom level to predict teacher-student interactions.

A handful of studies have applied audio and video analytics to entire classrooms in order to understand learning processes at a classroom level and develop an automated system of classroom assessment. For example, a series of studies by Raca and colleagues [10, 11] utilised videos of university lectures (students and teachers) to predict students' level of attention and engagement, while [12, 13] utilise audio recordings of classroom speech to automatically quantify classroom discourse patterns. Most of these studies focused exclusively on one stream of data, differing on our current study which harnesses both video and audio analytics to understand teacher-student interactions at the classroom level. A similar study [4] utilised audio, video, eye tracking, EEG measures to predict a teacher's actions and interaction level with the students (individual, group or whole class). However, this exploratory case study was carried out on one teacher, with low generalizability to other teachers, classroom environments and subjects.

III. Data and Challenges

The data is collected from 255 preschool classrooms in Singapore, by researchers of NIE (National Institute Of Education). Each classroom session is about 20 minutes, composed of 10 to 15 students. The recordings did not interfere with classroom instructions, enabling the capture of realistic conditions in a classroom. In these classrooms, a multitude of classroom activities, such as small team activities (children sitting around tables with the teacher walking around), and teacher-student discussion (students sitting around a circle/oval surrounding teacher) took place. Out of 255 recordings, 34 videos are selected for speaker diarization ground truth (GT), all of which are labelled independently by trained annotators. The audio with respect to speaker, speaking time and non-speech activities in class, are labelled, achieving an agreement of 80-95% between the 2 independent annotators. The speaker clusters, Teacher, Children, Overlap, Silence and Noise, are labelled as S0, S1, S2, S3, S4 respectively. Two professional observers, trained in CLASS scoring system, scored the overall classroom climate of each classroom, in accordance with rubrics outlined in the CLASS manual [1]. The observers looked at dimensions of positive affect, relationships, positive communication, and respect during teacher-student and student-student interactions when coding for positive climate. Similarly, when coding for negative climate, they assessed dimensions of negativity, punitive control and disrespect. The audio GT labels and their corresponding climate scores served as the GT for machine learning classifiers. We analysed 221 classrooms, which is about 74 hours of audio (20 min x 221 recordings).

Audio is recorded only by a single microphone worn by the teacher, capturing not only the speech of the teacher and children, but also ambient noise. (e.g., crying and rattling toys). The childrens' speech is occasionally misclassified as noise, since the single microphone failed to clearly pick up children speech with adequate fidelity and intelligibility: the reason is more often teacher is walking around and the children might be far away from the microphone and/or speaking softly. Most, if not all, state-of-the-art diarization algorithms were designed for non-moving microphone. Yet given the importance of speech and sound as a mode of interaction in the classroom, the resultant issues with the clarity of the audio on our recordings hindered our ability to adequately capture the teaching and learning processes taking place. On the other hand, the classroom videos are recorded with only a single camera set at

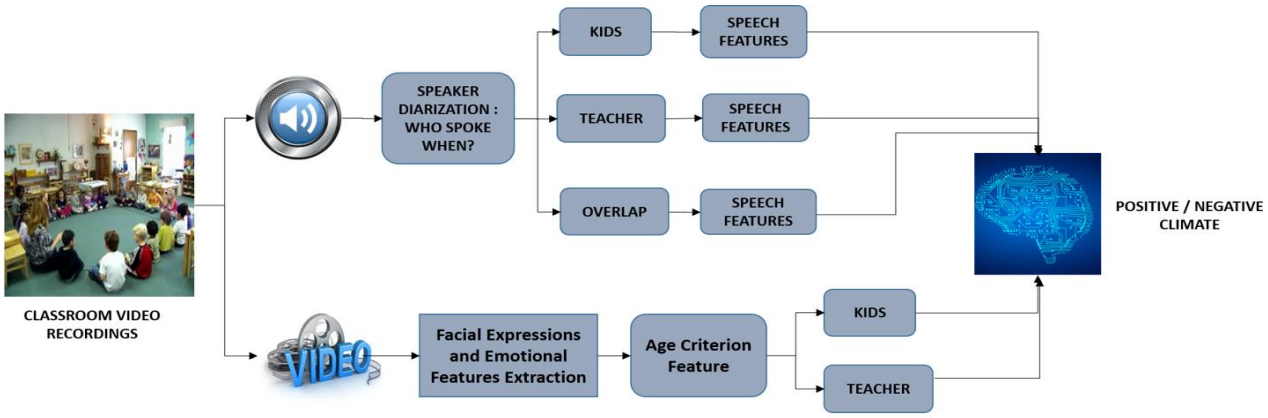

Figure 1: Designed Pipeline

an oblique angle, and the camera orientation varies for each classroom, providing various technical hurdles for the extraction of continuous and substantive features like facial expressions and emotions of all children and teachers.

Therefore, the research is novel in the following ways: First, unlike studies which investigate structured class settings, an ecological data set is considered here. Learning environments are more dynamic and less controlled than structured dialogs. Our project might be one of the first of this scale and ambition to target specifically the development of speech and video analytics tools for analysing learning activities; specifically, as a first step, we attempt inferring the classroom climate.

IV. Proposed Approach

This section elaborates on the proposed pipeline which consists of following steps: speaker diarization, audio and video feature extraction, sociometric analysis and training machine learning classifiers to predict the classroom climate (Fig. 1). In the following we explain each step-in detail.

(a) Speech Detection: First the system classifies speech and non-speech events from the recordings (speech detection), as it has been shown in [14]. Silence is removed from the audio by applying the thresholding technique and features such as energy, harmonic to noise ratio are applied to detect and remove non-speech segments [15].

(b) Speaker Group Diarization: In this step, the system identifies "who spoke when" (speaker group diarization). This approach has three steps: First, LIUM speaker diarization toolkit [16] identifies speakers by extracting acoustic features by segmenting, and clustering. LIUM's default settings yielded poor results as it was originally designed for broadcast news recordings and had difficulties to identify overlap speech of children and teacher. Therefore, for the second step, we applied a post-processing on the LIUM segments labelled as noise segments due to the overlap of speech and non-speech segments in the speech background. This enabled the retrieval of more speech segments (initially misclassified as noise), even in a noisy environment. Third, an automated classifier was developed to label each cluster segment as teacher (S0), children (S1), overlap (S2), silence (S3) and noise (S4), since LIUM provides general cluster names such as S100, S120. We compute the Euclidean distance between the low-level audio features, such as MFCCs (Mel-Frequency Cepstral Co-efficients), pitch etc, extracted from each LIUM cluster and the features extracted from the manually labelled GT segments (from the 34 audio recordings chosen for training the speaker diarization system).

Finally, we assign the label (S0, S1, S2, S3, S4) with the smallest Euclidean distance to each LIUM cluster. This step is similar as stated in [14].

(c) Feature Extraction: In this step, audio and video features are extracted. Audio feature extraction is same as stated in [14] and is briefly explained below.

A. Audio Features: We extracted conversational features from speech sequences (normalized speaker time, mean and standard deviation of speaker time, overlap counts and turn taking counts), and the means and standard deviations of 988 low-level audio features for each speech segment with the OpenSmile feature extraction tool [17]. Thus, a total of 5940 features (speech features for teacher, children and overlap segments) were obtained.

B. Video Features: This step is added to overcome the challenges in the approach proposed in [14]. We employed Affectiva [18], a feature extraction tool, to detect people in the video recordings and extract their facial expressions and emotions. Some of the extracted facial features include brow furrow, brow raise, jaw drop, smile, smirk etc. These facial features are mapped to emotions like joy, sadness, disgust, etc. These form the features for video analysis and have a metric range of 0 to 100. For each video, we utilised the age feature in Affectiva to distinguish between teacher's and children's facial and emotional features. We then computed the maximum, minimum, skewness, median, mean and kurtosis values for all features and across all time stamps for the teacher's and children's features separately. In total, 477 video features are obtained for teacher and children.

(d) Feature selection: As the number of features grows, the computational complexity will increase, and performance of the classifiers may degrade. Therefore, we applied Kruskal-Wallis test, followed by correlation-based feature selection algorithms to select the most relevant features, which are then fed as input

TABLE I.
RESULTS FOR CLIMATE DETECTION COMPARING AUDIO FEATURES, VIDEO FEATURES, COMBINATION OF ALL FEATURES, USING 10-FOLD CROSS-VALIDATION

	Audio Features only			Video features only			Audio and video features combined		
Classifier	Weighted average of Precision	Weighted average of Recall	Weighted average of F_1-score	Weighted average of Precision	Weighted average of Recall	Weighted average of F_1-score	Weighted average of Precision	Weighted average of Recall	Weighted average of F_1-score
Random-Forest	0.73	0.77	0.75	0.71	0.80	0.74	0.76	0.80	0.77
Gradient Boosting	0.75	0.78	0.76	0.74	0.80	0.75	0.75	0.80	0.77
Decision Tree	0.72	0.72	0.72	0.72	0.74	0.73	0.73	0.77	0.75
SVM-Linear Kernel	0.74	0.62	0.66	0.75	0.73	0.74	0.76	0.67	0.70
SVM-Gaussian Kernel	0.73	0.63	0.67	0.73	0.77	0.75	0.74	0.69	0.71
Adaboost	0.68	0.76	0.71	0.69	0.76	0.72	0.77	0.81	0.78
KNN	0.73	0.74	0.73	0.71	0.80	0.74	0.75	0.75	0.75
Logistic Regression	0.76	0.63	0.67	0.77	0.68	0.71	0.76	0.67	0.70
Gaussian Naive Bayes	0.76	0.75	0.75	0.73	0.54	0.58	0.75	0.78	0.76

to the classifiers for predicting the classroom climate. Feature selection methods employed in this step are the same as outlined in [14]. Correlation-based feature selection removes highly correlated features. The performance of classifiers is analysed by 10-fold cross-validation.

V. RESULTS AND DISCUSSION

In this section, we discuss the performance and analysis of audio features, video features and audio combined with video features. The dataset size is 221, of which 40 are negative climate classrooms and 181 are positive climate classrooms, resulting in class imbalance.

We trained and tested the speaker diarization system on 8 audio recordings (4 positive, 4 negative climate) which were chosen as gold standard set, to check performance of our speaker diazaration system. To obtain reliable evaluation results, we applied leave-one-out cross-validation (LOOCV); we trained the speaker diarization system on 7 audio recordings, tested it on the 8th audio recording, and repeated this procedure for all 8 recordings, and averaged the results. The speaker diarization system, as developed in [14], yields an average accuracy of 77% for teacher and 72% for children. This level of accuracy is reasonable as the students' speech was not captured with high fidelity. The accuracy for overlap is low, as LIUM is not designed to detect overlap. The overlap segments are often misclassified as teacher speech, since the voice of the teacher is typically recorded the loudest. However, overlap is relatively infrequent in our recordings and not critical in our analysis. Silence rarely occurs as the classes are quite active. Conversational and low-level audio features are extracted from the speech segments, which are derived from this speaker diarization system.

Similarly, video features are extracted from Affectiva and categorized into teacher and children features, based on the age feature in Affectiva. The performance of distinction between the teacher and children based on age criterion feature is manually tested, yielding an F_1-score of 70%.

For the purpose of manually testing the distinction between the teacher and children based on age criterion, we considered 9 classroom videos, for which the camera was oriented in the direction of the teacher for maximum duration of time. For each time instance output by the Affectiva tool the age criterion outcome is checked, if age was greater than 18 categorized as teacher, and if age was less than 18 categorized as student. This

outcome was verified by manually going through the videos, and checking at those time instances whether teachers and students age was correctly output by the Affectiva tool.

The Kruskal-Wallis test determines the discriminative features (with p-values < 0.005), in both the audio, video and combined analysis, for climate prediction. Spectral features like MFCC, pitch and LSP (Line Spectral Pairs) have the smallest p-values. Such features carry emotional information because their dependency on the tension of vocal folds [19]. The emotional content of speech is related to the acoustic characteristics of voice which can indicate emotions like joy, surprise, anger, or disgust [19]. The importance of these primary spectral features in emotion recognition has been established in [19] and [20]. Typically, we observe that low-level audio features are better for climate prediction than conversational features.

On the other hand, the Kruskal Wallis test for video features revealed that facial expression features such as the median, mean, maximum values of smirk, smile, and wink for teachers and children have smallest p-values; Similarly, facial emotional features such as maximum, mean, median and skewness values of surprise, sadness, valance, anger, disgust, and joy for both teachers and children have the smallest p-values. And these relevant video features are employed for climate prediction.

To determine if positive and negative classroom climate can be predicted based on audio, video and combined audio-video features, we trained and evaluated the following classifiers: Random-Forest, Gradient Boosting, C4.5 Decision Tree, SVM with Linear Kernel (SVML), SVM with Gaussian Kernel, Adaptive Boosting with C4.5 Decision Tree, k-Nearest Neighbours, Logistic Regression and Naïve Bayes and applied 10-fold cross-validation on 221 classroom recordings. An independent test set was not considered due to the limited data. The training labels are +1 (positive classroom climate) and -1 (negative classroom climate). In order to build the machine learning model and classify the classroom climate, we need to optimize the hyperparameters. Some of the hyperparameters were assigned aprior such as class weight, which helps to reduce the imbalance present in the dataset by giving higher weightage to minority class. It is calculated as:

$$\text{Class weight} = \left[\frac{\text{n\_samples}}{\text{n\_classess} \times \text{count of that particular class}}\right]$$

The decision boundary function is multiplied by the class weight in order to better classify the minority samples. Other hyperparameters are tuned by grid search.

All ensemble classifiers such as random-forest, gradient boosting and ada-boost classifiers are having number of base estimators as the hyperparameter. For SVM, kernel and the C-value are used as the hyperparameters, and KNN has number of neighbours as the hyperparameter. TABLE II. provides brief information about these hyperparameters for few of the classifiers which are crucial in building the machine learning model. Weighted average of precision, recall and F_1-score are employed as performance measures and are described with the abbreviation WAPM, for classifying the classroom recordings climate.

TABLE II.
CLASSIFIERS AND THEIR CORRESPONDING HYPERPARAMETERS USED IN MODEL ANALYSIS

Classifier	Hyperparameters
Random Forest Classifier	Number of base estimators: 10 Class weight: 'balanced'
SVM	Kernel: linear, radial basis function Class weight: 'balanced'
Logistic Regression	Class weight: 'balanced'
KNN	Number of neighbors:5

The weighted average score for a particular performance measure is defined as:

$$\text{WAPM} = (C1 \times P1 + C2 \times P2) \div (C1+C2)$$

WAPM: Weighted Average Performance Measure,
C1: Number of samples of positive class,
C2: Number of samples of negative class,
P1: Performance measure of positive class,
P2: Performance measure of negative class.

As mentioned before, the performance measures are precision, recall and F_1-score. Table I. summarizes performance results of audio features only, video features only and audio combined with video features. The baseline results for weighted average f-score in our case is 74%. Our model performance is slightly better than the baseline scores. For few of the classifiers such as ensemble classifiers, the positive climate which is the majority class has f-score of 80-90% as the average f-score among the 10-fold cross-validation, on the other hand negative climate which is the minority class has the average f-score of 30-40% on the 10-fold cross-validation. So the overall weighted f-score varies between 70 to 78% for these classifiers which is as described in TABLE I.

We observe that, with the audio features only, Logistic Regression and Naïve Bayes yield the best weighted average precision of 76%, while Gradient Boosting leads to the best weighted average recall of 78% and weighted F_1-score of 76%. On the other hand, when employing only the video features for climate prediction, the weighted average precision did not improve significantly in comparison with that of the audio features. However, for most classifiers, the weighted recall and F_1-scores are higher when video features only are employed. Finally, when combining audio and video features for climate prediction, the best weighted F_1-scores are obtained mostly for all classifiers with the exception of Logistic Regression, SVMG and SVML. These results suggest that for climate prediction, both audio and video features are crucial and provide complementary information for the inference of climate scores, illustrating the potential of leveraging audio-visual features for enhanced climate prediction.

VI. Conclusion And Future work

In this paper, an automated way of classifying between positive and negative classroom climate is accomplished, in order to provide instant feedback to teachers regarding their behaviour during the classroom period. Our pipeline architectures describe how speech and video analysis tools with machine learning approaches could help in providing such instant feedbacks about the classroom climate. As part of the speech analysis, we extracted conversational and prosodic features of teachers and children after speaker diarization is performed. Similarly, in the video analysis, the extraction of facial and emotional features of the teachers and children is performed. Finally, an automated model with these features for classroom climate prediction is developed. We observe that combining audio and video features leads to an improvement in the performance metrics. Our results underscore the possibility of predicting classroom climate by employing low-level audio features and facial expressions, even from the recordings captured by a single microphone worn by the teacher in noisy classrooms. Also, the results indicate that video features can also help in a better detection of classroom climate, even with the limited facial information that was captured by a video recorder, set in an oblique angle.

Future work includes exploring the other audio-video analytic techniques by applying different feature transformation techniques on speech and video extracted features. Also, performing sentiment analysis on the textual data of the classrooms by converting speech to textual data, and also identifying different kinds of intents of teachers and students from the textual data can be explored. In the future we would like to investigate the silences in recordings, and their impact on the climate classification. Further, on the video front, identifying the different postures of teachers and students which will intern lead to gesture recognition in the classroom video recordings for teachers and students can be investigated. In an attempt to improve the overall performance, more data samples will be collected for negative classroom climate in order to make the dataset more balanced. Also, we would like to apply different statistical learning techniques and feature engineering techniques to validate our system thoroughly. Finally, work towards building and validating an automated assessment system will be carried out to generate an effective timely feedback from the classroom recordings which will be deployed to real-world classrooms.

ACKNOWLEDGEMENTS

This project is supported by a grant from Centre for Research and Development in Learning (CRADLE@NTU).

REFERENCES

[1] R.C., Pianta, K. M. La Paro, and B. K. Hamre, Classroom Assessment Scoring System™: Manual K-3. Classroom Assessment Scoring System™: Manual K-3. 2008, Baltimore, MD, US: Paul H Brookes Publishing. xi, 112-xi, 112.

[2] M., Worsley and P., Blikstein, "Leveraging Multimodal Learning Analytics to Differentiate Student Learning Strategies," In Proceedings of the Fifth International Conference on Learning Analytics and Knowledge, New York, NY, USA: ACM, pp. 360–367, 2015

[3] P., Blikstein and M., Worsley, "Multimodal Learning Analytics and Education Data Mining: Using Computational Technologies to Measure Complex Learning Tasks. Journal of Learning Analytics, vol 3, pp. 220-238, 2016.

[4] L. P., Prieto, K., Sharma, P., Dillenbourg, and M., Jesús, "Teaching analytics: Towards automatic extraction of orchestration graphs using wear-able sensors," in Proceedings of the Sixth International Conference on Learning Analytics and Knowledge—LAK '16, 2016, pp.148–157.

[5] D. Spikol, Daniel & E., Ruffaldi, Emanuele & G., Dabisias, and M., Cukurova, "Supervised machine learning in multimodal learning analytics for estimating success in project-based learning," Journal of Computer Assisted Learning, 2018

[6] X. Ochoa, K., Chiluiza, G., Méndez, G., Luzardo, B., Guamán, and J. Castells, "Expertise estimation based on simple multimodal features," In Proceedings of the 15th ACM International Conference on Multimodal Iter-action (ICMI '13), ACM Press, New York, 2013, pp. 583–590.

[7] S., Scherer, N., Weibel, L. P., Morency, and S., Oviatt, "Multimodal prediction of expertise and leadership in learning groups," in Proceedings of the 1st International Workshop on Multimodal Learning Analytics (MLA '12), 2012, pp. 1–8.

[8] A. Andrade, "Understanding Student Learning Trajectories Using Multimodal Learning Analytics within an Embodied-Interaction Learning Environment," in International Learning Analytics and Knowledge Conference, 2017.

[9] C, Smith, B, King, and D., Gonzalez, "Using Multimodal Learning Analytics to Identify Patterns of Interactions in a Body-Based Mathematics Activity," Journal of Interactive Learning Research, vol. 27, issue 4, pp.355-379, 2016.

[10] M., Raca, R., Tormey, and P. Dillenbourg, "Sleepers' lag-study on motion and attention," In Proceedings of the Fourth International Conference on Learning Analytics And Knowledge, 2014, ACM, pp. 36-43.

[11] M., Raca, and P., Dillenbourg, "Holistic Analysis of the Classroom," in Proceedings of the 2014 ACM Workshop on Multimodal Learning Analytics Workshop and Grand Challenge - MLA '14, 2014, pp. 13–20.

[12] K. D' Mello et al., "Multimodal Capture of Teacher-Student Interactions for Automated Dialogic Analysis in Live Classrooms, in Proceedings of the 2015 ACM on International Conference on Multimodal Interaction. 2015, ACM: Seattle, Washington, USA. pp. 557-566.

[13] P., Donnelly et al., Words matter: automatic detection of teacher questions in live classroom discourse using linguistics, acoustics, and context, pp. 218-227, 2017

[14] James, Anusha et al, "Automated classification of classroom climate by audio analysis", in proceedings of Ninth International Workshop on Spoken Dialogue Systems Technology (IWSDS 2018).

[15] T. Giannakopoulos and A. Pikrakis, "Introduction to Audio Analysis: A MATLAB Approach," 2014, Oxford: Academic Press.

[16] S. Meignier, and T. Merlin. "Lium Spkdiarization: an Open Source Toolkit for Diarization", 2009.

[17] Opensmile book. http://www.audeering.com/research-and-opensource/files/openSMILE-book-latest.pdf

[18] D. McDuff, A. Mahmoud, M. Mavadati, M. Amr, J. Turcot, R. Kaliouby, "AFFDEX SDK: A Cross-Platform Real-Time Multi-Face Expression Recognition Toolkit," in Proceedings of the CHI Conference Extended Abstracts on Human Factors in Computing Systems, 2016, pp. 3723-3724.

[19] M. C., Sezgin, B. Gunsel, and G. K. Kurt, "Perceptual audio features for emotion detection". EURASIP Journal on Audio, Speech, and Music Processing, vol. 1, 2012.

[20] H., Gunes, et al. "Emotion representation, analysis and synthesis in continuous space: A survey," in Face and Gesture, 2011

[21] N.Bosch, K.D'mello, J.Ocumpaucgh, R.Baker, V.Shute, "Using Video to Automatically Detect Learner Affect in Computer-enabled Classrooms" in Proceedings of the ACM Transcations on Interactive Intelligent Systems, 2016, pp. 6 (2).