Automated Classification of Classroom Climate by Audio Analysis

Anusha James*, Chua Yi Han Victoria*, Tomasz Maszczyk*, Ana Moreno Núñez†, Rebecca Bull†, Kerry Lee†, Justin Dauwels*

*School of Electrical and Electronic Engineering (EEE), Nanyang Technological University,
†National Institute of Education, Singapore

Abstract: While in training, teachers are often given feedback about their teaching style by experts who observe the classroom. Trained observer coding of classroom such as the Classroom Assessment Scoring System (CLASS) provides valuable feedback to teachers, but the turnover time for observing and coding makes it hard to generate instant feedback. We aim to design technological platforms that analyze real-life data in learning environments, and generate automatic objective assessments in real-time. To this end, we adopted state-of- the-art speech processing technologies and conducted trials in real-life teaching environments. Although much attention has been devoted to speech processing for numerous applications, few researchers have attempted to apply speech processing for analyzing activities in classrooms. To address this shortcoming, we developed speech processing algorithms that detect speakers and social behavior from audio recordings in classrooms. Specifically, we aim to infer the climate in the classroom from non-verbal speech cues. We extract non-verbal speech cues and low-level audio features from speech segments and we train classifiers based on those cues. We were able to distinguish between positive and negative CLASS climate scores with 70-80% accuracy (estimated by leave-one-out crossvalidation). The results indicate the potential of predicting classroom climate automatically from audio recordings.

1. INTRODUCTION

CLASS [1] assesses the interactional quality between students and teachers in preschool classrooms and two of its ten subscales are related to the climate in the classrooms. Classroom climate is the overall emotional tone of the class observed from the warmth and respect in interactions. Students in a positive climate class are enthusiastic and happy, whereas negative climate reflects anger or aggression. Establishing a positive climate in class is important for effective teaching [1]. It is tedious and time-consuming to manually observe and code CLASS scores. There is a need to explore indicators that are easily and quickly extracted and correlate with CLASS climate dimensions. As classroom discourse is an important indicator of teacher instruction and classroom climate, we explore the idea here to analyze teaching practices automatically by means of speech processing technologies. Specifically, we have designed a pipeline to infer the classroom climate automatically. The processing contains the following steps: speaker diarization, sociometric analysis and machine

learning to audio recordings of preschool classrooms, which posed a few technical challenges. In this paper, we describe each step of the processing pipeline, and show numerical results both for speaker diarization and estimation of the classroom climate. Overall, we achieved a classification accuracy for classroom climate between 70% and 80%. These results are promising, since preschool classrooms tend to be noisy, and the audio was only recorded by a microphone worn by the teacher.

The rest of the paper is structured as follows. In section 2, we briefly summarize similar work, while in Section 3, we discuss about the data and challenges that motivated this framework. In Section 4, we elaborate on the algorithms developed, whereas in Section 5 we discuss the results and findings, followed by conclusions in Section 6.

2. RELATED WORK

In a related study, Blanchard et al. developed a pipeline that automates Nystraand's CLASS coding scheme for dialogic instruction [2]. Natural language processing, utterance timing, and acoustic features were exploited to automatically detect teacher's questions ($F_1 = 0.69$) [3] and classify instructional segments ($F_1 = 0.60$), such as question-and-answer [4] in middle-school classroom audio. Most of Blanchard's studies on classroom discourse classification are centered around automated analysis of the teacher speech signals only, whereas we attempt to analyze both the teacher and children speech. Similarly, Wang et al. developed tools to automatically segment recordings from the Learning Environment Analysis system (LENA) [5, 6] into teacher speech, student speech, overlap, silence and discussion. Next they classified the teaching behaviors by leveraging on conversational features (i.e., length of time student/teacher spoke/discussion). When given feedback on these features, teachers promoted desirable teaching behaviors by giving more time to students to speak, and by making more time for discussions [5]. Lastly, a large-scale study, spanning 1720 hours of audio from college science courses [7], applied decibel analysis to automatically classify between lecture and non-lecture activities. The aforementioned studies all rely on speech features to automatically predict types of classroom activities, while in our study we aim to predict classroom assessment scores; specifically, as a first step, we attempt inferring the classroom climate.

3. DATA AND CHALLENGES

The data is collected from 92 classrooms in multipole preschools in Singapore by researchers of the National Institute of Education (NIE). The recordings typically last 20 minutes, comprised 10 to 15 students, and features different classroom activities, such as small team activities (children sitting at tables with teacher walking around), and teacher-student discussions (students sitting around teacher). Each video was captured by one stationary camera, and the audio is recorded by a microphone worn by the teacher.

For the climate labels, two independent annotators scored the overall classroom climate in each video according to the rubrics outlined in the CLASS manual [1].

Annotators 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.

In this paper, we analyzed only the microphone audio recordings. Out of 92 videos, we selected 12 videos to label audio ground truth (GT). Annotators were trained to label the audio w.r.t. speakers, speaking time, and non-speech activities. The nomenclature is as follows: Teacher (S0), Children (S1), Overlap - when teacher and children speak together (S2), Silence (S3), and Noise (S4). The two annotators achieved an agreement of 80-95%, while labelling independently. We analyzed 80 audio recordings, which is about 27 hours of audio (20 min x 80 audio recordings).

Building a robust speaker diarization system in this setting is a challenging task due to varying classroom settings. The audio is only recorded by a single microphone worn by the teacher, capturing not only teacher and children speech, but also background noise (e.g., crying and feet stamping). The speech of the children was not always captured with adequate fidelity and intelligibility, since the teacher is often walking around and the children might be far away from the microphone and might be speaking softly; as a result, the children speech is occasionally misclassified as noise. Most, if not all, state-of-the-art diarization algorithms were designed for single non-moving microphone. 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. Second, the ground truth CLASS scores are coded w.r.t. visual and audio information, while we only used audio information to predict CLASS climate scores.

4. PROPOSED APPROACH

In this section, we present the proposed speech processing pipeline. First the system classifies speech and non-speech events from the recordings (speech detection), it identifies "who spoke when" (speaker group diarization). Next it extracts conversational and low-level features of teacher, students, overlap, silence and noise segments and at last, it infers from those features the classroom climate. In the following we explain each step in detail.

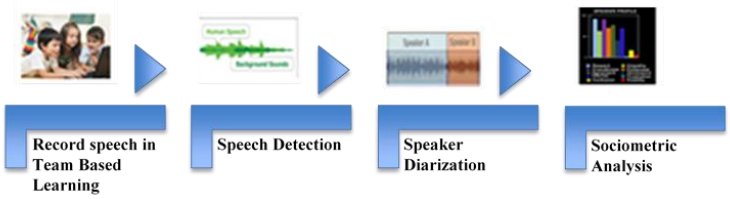

Fig.1: Block diagram of the proposed approach.

4.1 Speech Detection: It is advantageous to first separate speech from non-speech acoustic activity like music, background noise, noise made using toys, etc [8]. We im-

plemented the following approach; Silence is removed from audio by means of thresholding, and next the resulting audio is passed to a non-speech event detector. Some of the input features for this detector include energy and harmonic ratio to detect non-speech [9]. As a result, only the speech segments are retained.

4.2 Speaker Group Diarization: In this step, we identify different group of speakers present, and attribute each speaker with their spoken speech segments. We applied the LIUM toolkit [10], which performs diarization by extracting acoustic features, segmenting, and clustering. LIUM's default settings yielded poor results as it was originally designed for broadcast news recordings, and had difficulties parsing overlap speech of children and teacher.

4.3 Automated Speaker Labelling: LIUM provides general cluster names such as S100, S120, but for our purposes, these segments need to be labelled specifically as teacher (S0), children (S1), overlap (S2), silence (S3) and noise (S4). To this end, we developed an automated classifier to label each cluster. We compute the Euclidean distance between, on the one hand, the low-level audio features, such as MFCCs, pitch etc, extracted from each LIUM cluster and on the other hand, the features extracted from the manually labeled GT segments (from 12 audio recordings chosen for training the speaker diarization system). We then assign the label (S0, S1, S2, S3, S4) with the smallest Euclidean distance to each LIUM cluster.

4.4 Feature extraction and selection: We computed conversational features from speech sequences, e.g., normalized speaker time, speaker duration, and in addition, we extracted 988 low-level audio features for each speech segment by OpenSmile [11]. A total of 5940 features for teacher, children and overlap segments were obtained (see Fig 2). As the number of features grows, the computational complexity will increase and performance of the classifiers may degrade. Therefore, we applied Kruskal-Wallis test, and correlation based feature selection algorithms to select the most relevant features, which we feed into classifiers for predicting the classroom climate. We assess these classifiers by 10-fold cross validation.

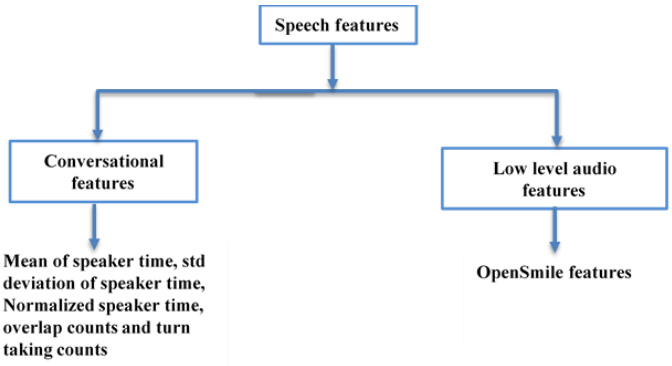

Fig 2: Features for climate prediction.

5. DISCUSSION / ANALYSIS AND RESULTS

5.1 Speaker Group Diarization: We trained and tested the speaker diarization system on 8 audio recordings (4 positive, 4 negative climate). To obtain reliable evaluation results, we applied 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 proposed system yields an average accuracy of 77% for teacher and 72% for children (Table 1a). 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 (see Table 1c). Silence rarely occurs as the classes are quite active.

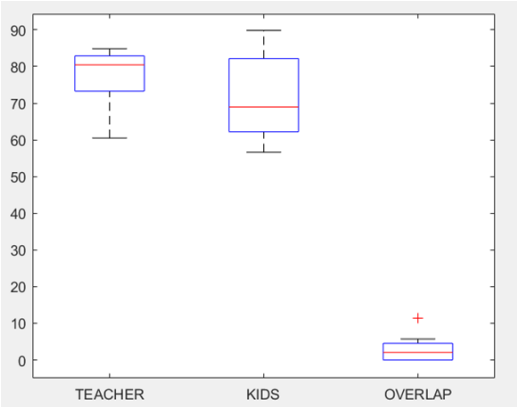

Fig 3. Boxplot of accuracy of speaker diarization of 8 classroom conversations.

(a) **Average of Normalized Confusion matrices (%)**			
True \ Estimate	**Teacher**	**Children**	**Overlap**
Teacher	**77.3**	**21**	**1.7**
Children	**23.7**	**71.6**	**4.7**
Overlap	**75.4**	**21.6**	**3.1**
(b) **Absolute Confusion matrices (in sec)**			
Teacher	**2773.4**	**741.2**	**61.4**
Children	**618.4**	**1885.6**	**142.6**
Overlap	**483**	**169.2**	**29.9**
(c) **Confusion matrix normalized by total time (%)**			
Teacher	**40.2**	**10.7**	**0.9**
Children	**9.0**	**27.3**	**2.1**
Overlap	**7.0**	**2.5**	**0.4**

Table 1. Confusion matrices showing the actual / hypothesized speakers association for 8 classroom conversations.

5.2 Climate Prediction: Conversational features and low-level audio features are extracted from the individual speaker segments. The Kruskal-Wallis test is applied on 5940 features to refine the discriminative features (with p-values < 0.005) for climate prediction (Fig 4a). 116 features (2%) satisfy this condition, of which 37 (32%), 36 (31%), and 43 (37%) features are associated with teacher, children and overlap respectively (Fig 4b). Spectral features like MFCC, pitch and LSP have the smallest p-values. Such features carry emotional information because of its dependency on the tension of vocal folds [12]. The emotional content of speech is related to the acoustic characteristics of voice which can indicate emotions like joy, surprise, anger, or disgust [12]. The importance of these primary spectral features in emotion recognition has been established in [12] and [13]. Typically, low-level audio features are better capable of climate detection than conversational features.

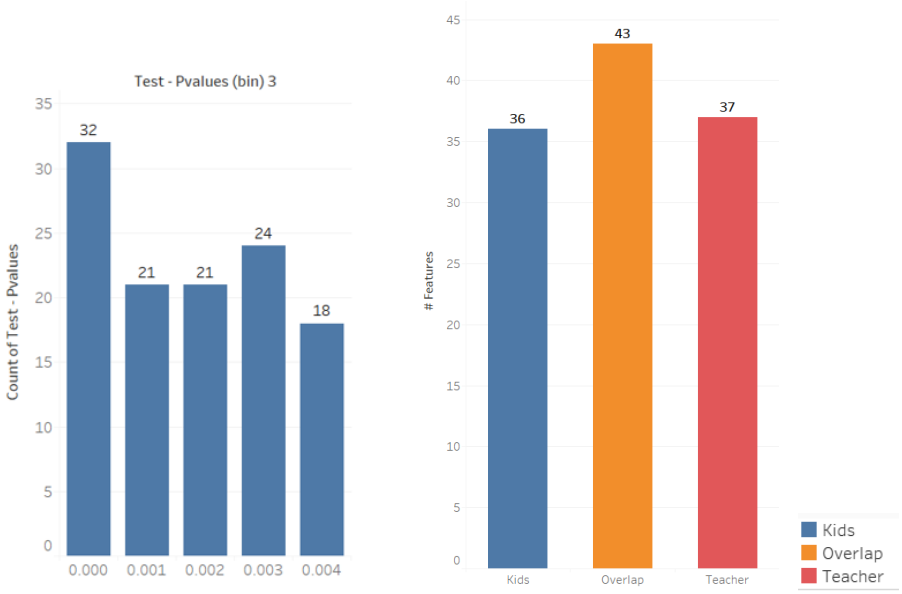

Fig 4. Statistics of salient features. a) Distribution of p-values; b) Number of salient features associated with teacher and children speech, and overlap segments.

Classifier	Crossvalidation accuracy
Linear SVM (Support Vector Machine)	0.71
Radial SVM (SVMG)	0.73
Logistic Regression(LR)	0.78
K Nearest Neighbors (kNN)	0.73
Decision Tree (DT)	0.60
AdaBoost Classifier (AB)	0.70
Random Forest Classifier(RF)	0.74
Naïve Bayes(NB)	0.79
Multilayer Perceptron (MLP)	0.80

Table 2: Mean of classification accuracy of 10 folds for different classifiers.

We reduced the feature set further by means of the random forest classifier approach [14] and by a correlation based approach [15]. We trained 9 classifiers with this reduced feature set and applied 10-fold CV on 80 recordings. These 80 recordings are different from the 12 recordings used to train the speaker diarization system. The training labels are “Positive climate” (+1) and “Negative climate” (-1). The accuracy of the classifier for the 10 folds is shown in Figure 5 whereas the corresponding mean accuracy is shown in Table 2. Overall, the classifiers yielded accuracies from 70%-80%. Multilayer Perceptron (MLP) yielded the best accuracy of 80%. We consider these results promising due to the challenging nature of the preschool classroom context. And also the entire framework relies only on the audio captured using teacher’s microphone.

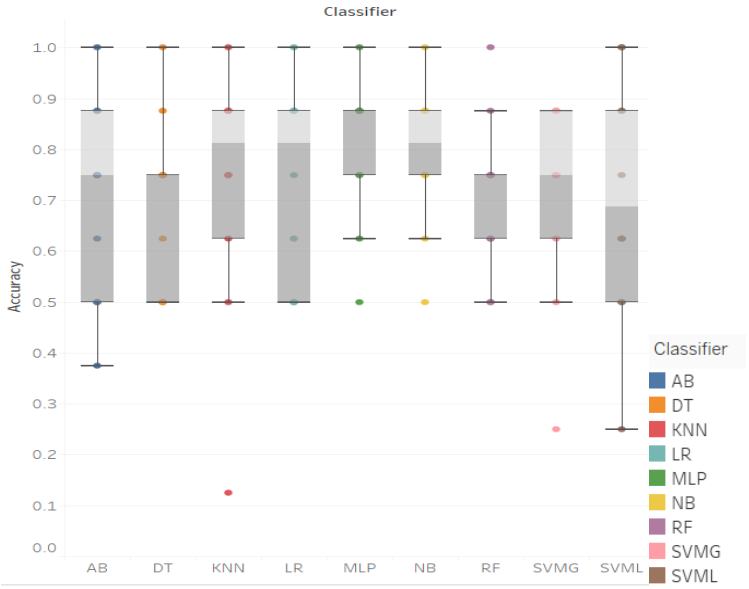

Fig 5. Boxplot of accuracy of classifiers in 10 folds.

6. CONCLUSION

The experimental results indicate that prediction of classroom climate might be possible from non-speech verbal and prosodic cues, even from recordings captured by a single microphone worn by the teacher in noisy classrooms. The ability to predict climate scores from low-level indicators could be useful in the development of automated feedback tools for aiding professional teacher development. More testing is warranted to investigate how such a system can be adapted to suit the needs of professional development and classroom assessment. In future work, we will investigate video features and further validate and evaluate our system on a larger and more diverse dataset across more teachers, classroom sessions and class activities.

REFERENCES

1. Pianta, R.C., 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. Nystrand, M., CLASS 4.0 user's manual. 2004.
3. J. Donnelly, P., et al., Words matter: automatic detection of teacher questions in live classroom discourse using linguistics, acoustics, and context. 2017. 218-227.
4. Donnelly, P.J., et al., Multi-sensor modeling of teacher instructional segments in live classrooms, in Proceedings of the 18th ACM International Conference on Multimodal Interaction. 2016, ACM: Tokyo, Japan. p. 177-184.
5. Wang, Z., K. Miller, and K. Cortina, Using the LENA in teacher training: Promoting student involvement through automated feedback. Vol. 4. 2013.
6. Wang, Z., et al., Automatic classification of activities in classroom discourse. Computers & Education, 2014. 78: p. 115-123.
7. T. Owens, M., et al., Classroom sound can be used to classify teaching practices in college science courses. Vol. 114. 2017. 201618693.
8. Anguera, X., et al., Speaker Diarization: A Review of Recent Research. IEEE Transactions on Audio, Speech, and Language Processing, 2012. 20(2): p. 356-370.
9. Giannakopoulos, T. and A. Pikrakis, Introduction to Audio Analysis: A MATLAB Approach. 2014, Oxford: Academic Press.
10. Meignier, S. and T. Merlin. Lium Spkdiarization: an Open Source Toolkit for Diarization. 2009.
11. Opensmile book. http://www.audeering.com/research-and-open-source/files/openSMILE-book-latest.pdf.
12. Sezgin, M.C., B. Gunsel, and G.K. Kurt, Perceptual audio features for emotion detection. EURASIP Journal on Audio, Speech, and Music Processing, 2012. 2012(1): p. 16.
13. Gunes, H., et al. Emotion representation, analysis and synthesis in continuous space: A survey. in Face and Gesture 2011. 2011.
14. Breiman, L. Machine Learning (2001) 45: 5. https://doi.org/10.1023/A:1010933404324
15. Hall, Mark Andrew. "Correlation-based feature selection for machine learning." (1999).