Analysis of classroom interaction using speaker diarization and discourse features from audio recordings

Oscar Canovas and Felix J. Garcia

Department of Computer Engineering and Technology,
University of Murcia, Murcia 30100, Spain
{ocanovas,fgarcia}@um.es

Abstract. In this paper we present the rationale, and some initial results, of an automated system for classroom analysis which is based on speaker diarization techniques and non-verbal discourse features extracted from audio recordings. We have employed several Machine Learning algorithms and audio processing methods with classroom recordings related to several undergraduate courses. After determining the identity of the speaker in a recorded class, we can distinguish whether the speaker is a teacher, a student, there are multiple speakers at the same time, or silence. An important contribution of our work is that, from that information, we derive several non-verbal features that can be used to describe patterns. Our preliminary results show that it is possible to extract valuable information using data visualization. As we show, different teachers and teaching methods generate identifiable patterns, that might be used to analyze, for example, which methodologies and teaching styles provide higher levels of interaction or participation.

Keywords: Audio analysis, teaching practices, data visualization

1 Introduction and motivation

Many educational research findings have shown that the successful student interaction in a classroom, that is, sharing ideas with the teacher or participating in peer discussions, is correlated with high quality learning [7]. However, student-teacher relationships and interactions are complex and there is not a one-size-fits-all approach. We believe that the quality of interactions between teachers and students can be measured through standardized observation methods providing teachers with data about relevant features from classroom interactions. It has been proven that by analyzing their own lessons teachers can improve their student's academic achievement [11].

The main purpose of this work in progress is to present the rationale, and some initial results, of an automated system for analysis of classroom profiles which is based on speaker diarization techniques and discourse features from audio recordings. Identifying the identity of the speaker in a recorded class, we can distinguish whether the speaker is a teacher, a student, there are multiple speakers at the same time, or silence. An important contribution from our work is that, from that information, we derive several discourse features that can be used not only for classification purposes, but they are also informative because they describe patterns and measure levels of interaction. Therefore, our preliminary results show that it is possible to provide timely generation of a classroom profile based on data visualization to help teachers to analyze and improve their classroom activities.

We have developed and integrated data analysis techniques based on machine learning to diarize the audios, that is, to automate the annotation of when each speaker (or type of speaker) is talking. This diarization information is then used to extract features that describe specific aspects of the classroom activities, providing data about the participation rate, turn taking, pauses, or multiple speakers. These features will be used in future works to perform an automatic classification of the different activities and teaching methods. At his moment, they are used to create a classroom profile, using data visualization techniques.

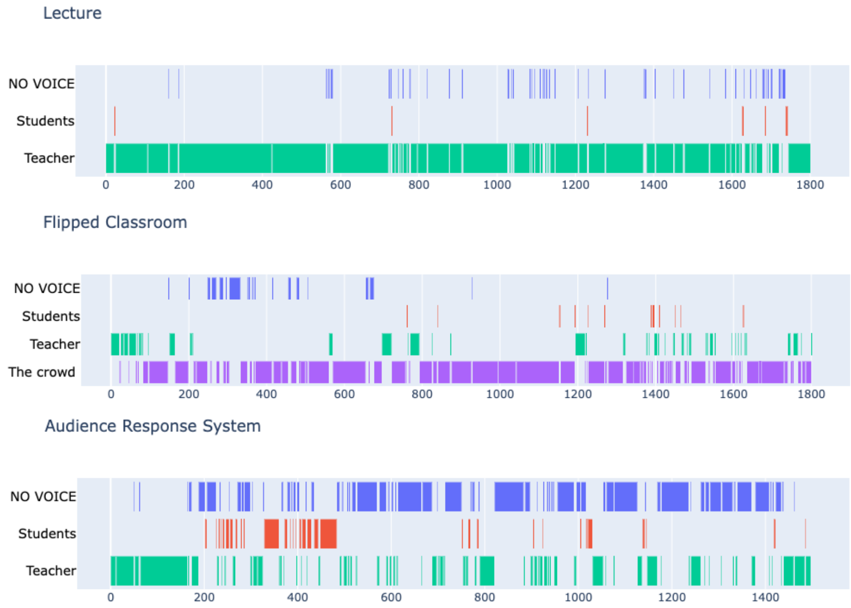


Fig. 1: Timelines of different audio patterns for different teaching methods

Our preliminary results are very promising about the suitability of the system to provide timely feedback for the teachers. After analyzing audios from different classroom activities, such as lectures, problem solving, flipped classrooms, or the use of audience response systems, we have confirmed that the results from the diarization process show very different patterns for each type of activity, as Figure 1 shows. We have also observed contrasting patterns for the different teachers performing the same type of activity. Finally, we have confirmed with the participating teachers that it is possible to grasp meaningful insights about interaction and teaching styles from the information displayed by our system.

2 Related work

In recent years, we can find several works analyzing the classroom climate and discourse. As it is our case, some of them employ audio recordings and machine or deep learning techniques to analyze different teaching practices and styles. Those recordings can be analyzed using non-verbal features or natural language processing techniques [10]. This latter option is out of the scope of our preliminary work, so we will focus on works based on non-verbal features.

Usually, the first step is to label the audio recordings to identify ”who spoke when”. There are multiple approaches to address this task, from classical methods to advanced neural networks [9]. Some previous methods for classroom audio analysis relied on participants to wear individual microphones [14] and the use of Language ENvironment Analysis (LENA) to achieve automated unsupervised classification of class activities. However, James et al. [3], as it is our case, adopted recent state-of-the-art speech processing technologies, and non-intrusive set-ups, to detect speakers and to infer the climate in the classroom from non-verbal speech cues.

Different research works have used machine learning models to detect teaching practices using spectral audio features, paying special attention to the role of the teacher [2, 12] or classifying active learning tasks [5, 8, 13]. In general, these works are specially focused on the classification accuracy of the employed techniques and they do not provide discourse features that can be used as descriptive and informational data.

Some of the discourse features in Section 5 were adapted from those presented in [1, 4] for group meeting analysis, since some of the turn-taking features, participation rates, or silence ratios are also suitable for the teaching analysis.

3 Context

In this work we analyze the audios recorded from two courses (Computer Networks and Computing Foundations) of a bachelor’s degree in Computer Science. In particular, four teachers from our university have weekly recorded their classes. The recording setup was unobtrusive since it consisted of placing a handheld digital recorder (TASCAM DR-07X) located at least 1.5 meters apart from the teacher and the front row students. Consequently, we ended up with a dataset of twelve recordings for each teacher, with lengths ranging from 90 to 125 minutes per audio file, and registering up to four different teaching methods (lecture, problem solving, audience response systems, and flipped classrooms). The dataset was completed with a weekly survey submitted by the teachers about the student participation level per class and a final survey from the students about teacher’s instructional support. All the information was collected with the approval of the teachers, students and the Institutional Review Board.

4 Audio pipeline

This section provides details about the designed pipeline shown in Figure 2. First, an initial audio processing extracts low level features from the audio recordings. There is a thresholding method to separate silence from voiced segments, and, using small audio frames, we compute Mel Frequency Cepstral Coefficients (MFCCs) [6], pitch and average energy for each voiced frame. These features are widely used in the scientific literature as input for speaker diarization techniques.

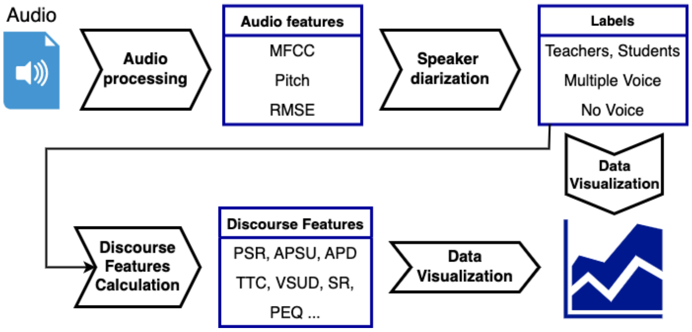


Fig. 2: Pipeline of processing stages and data elements

The second stage is to determine ”who spoke when”, that is, to label every second of each recording with the corresponding speaker (or silence). In our preliminary prototype, we do not distinguish which particular student is actually speaking, and all the intervening students are labeled just as ”Students”. We also annotate if there is silence or multiple voices. This diarization process is based on machine learning techniques making use of Support Vector Machines, and it is a semi-supervised process as we have to train the system with some audio fragments that are tagged as ”Teacher”, ”Students” or ”The crowd”. This method only requires 30 seconds of audio, for each category, to provide a classification accuracy around 91% with other test audios. Consequently, we obtain an time-ordered sequence of labels that will be used in the next stage of the pipeline but it can be also be visualized using timelines, as Figure 1 showed.

The third step is the calculation of non-verbal discourse features from the sequence of labels. As we detail in Section 5, we compute participant speaking rates, average pause duration, turn taking counts, etc.

Finally, as we will show in Section 6 we provide several data visualization alternatives to display the computed features and also the results from the diarization process. The main rationale behind those graphs is to provide straightforward and valuable information to the teachers in order to analyze some classroom activities, such as levels of interaction or teaching styles.

5 Features for audio analysis

As we mentioned previously, considering some previous works [1, 4], we have defined several non-verbal features extracted from audio recordings that are useful for the characterization of discourses. Specifically, we define two discourse features per role (teacher, students, the crowd, or silence):

- **Participant Speaking Ratio (PSR)**. It measures the ratio of participation of each role during the classroom recording.
- **Average Participant Speaking Utterance Duration (APSUD)**. Measured in seconds, it is the average duration of the utterances of each participant.

Additionally we have also defined five global discourse features:

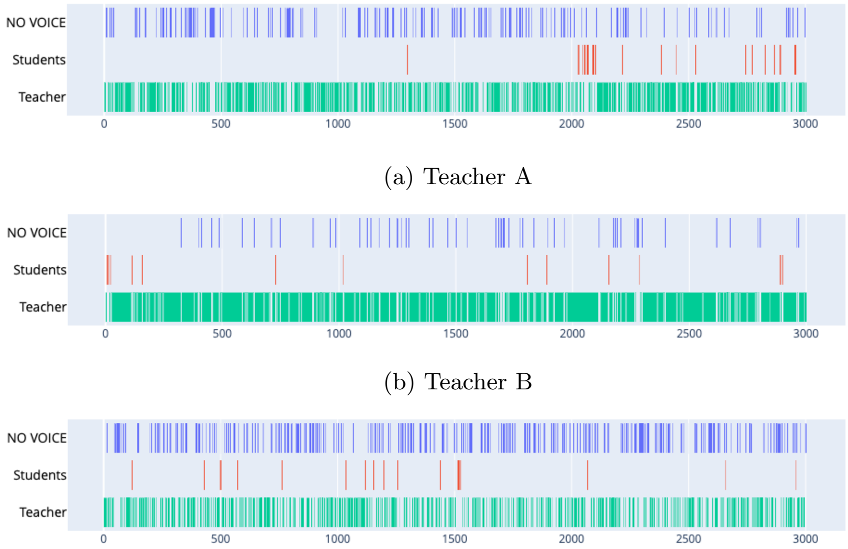
- **Average Pause Duration (APD)**. It measures how long is the average silence interval between utterances of the same participant.
- **Participation equality (PEQ)**. It is an indicator of how balanced is the participation of the different roles and it is calculated as shown in [4]. Values close to 1 denote equal participation.
- **Turn taking count (TTC)**. It represents how many turn changes occurred in the dialogue between students and teacher.
- **Silence Ratio (SR)**. It measures the ratio of silence during the classroom recording.
- **Very Short Utterances Ratio (VSUR)**. It shows the ratio of very short speech utterances (less than 2 seconds) over the total.

6 Insights from data visualization

One of the main goals of this work is to provide valuable visualization for the data derived from the audio processing pipeline. Our prototype uses different graphs, such as timelines, bars, and distribution functions, to gain some knowledge about the classroom activities.

After an initial analysis of the recordings obtained from the set-up described in Section 3, we observed interesting patterns that show the differences related to different teaching methods. For example, when we apply the diarization process to recordings related to different methodologies (for the same teacher) we perceive distinguishable patterns for each of them in the timelines, as Figure 1 showed. Lectures provide lower levels of interaction (the students' participation is scarce), flipped classrooms tend to be more participatory and noisy (in the example the class was solving problems in small groups) and the use of audio response systems involves alternating periods of silence, explanations, and questions.

We have also analyzed whether the timeline representation allow us to detect different discourse patterns for different teachers who are lecturing the same contents in their respective classrooms. As Figure 3 shows, we observe some



(a) Teacher A

(b) Teacher B

(c) Teacher C

Fig. 3: Timelines of different teachers but the same lecture contents

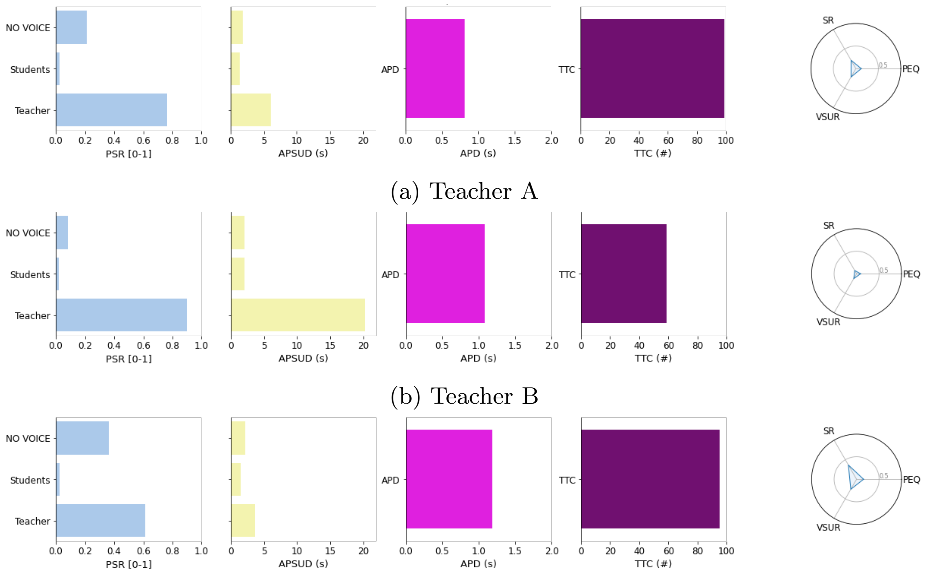
clear differences in relation to the amount of ”NO VOICE” or silence patterns, that is, the number of pauses made by each teacher.

However, these differences are more clearly visible when we compute the non-verbal discourse features for each audio and we visualize the results, as Figure 4 shows. Teachers can analyze the level of participation of the students (in this case is less than 4% for all the classes), their participation rates or the average utterance duration (with notorious differences between teachers), turn taking counts, participation equality, etc.

Even if the teachers do not have access to the data related to other teachers, they can find valuable information about their own teaching style and levels of interaction after each class. In fact, our prototype also visualizes data about the statistical distribution and temporal evolution of each feature. Figure 5 shows, for example, three different graphs (cumulative distribution function, histogram and time-ordered) for one feature, teacher participation rate, along the whole course. In this case, the histogram shows that teacher participation tends to be bimodal, depending on the teaching method (higher for lectures and lower for flipped classroom).

7 Conclusions and future work

In this work in progress, we have successfully developed a system for visualizing different features, extracted from audio recordings, that can be used to provide



(a) Teacher A

(b) Teacher B

(c) Teacher C

Fig. 4: Discourse features for different teachers but the same lecture contents

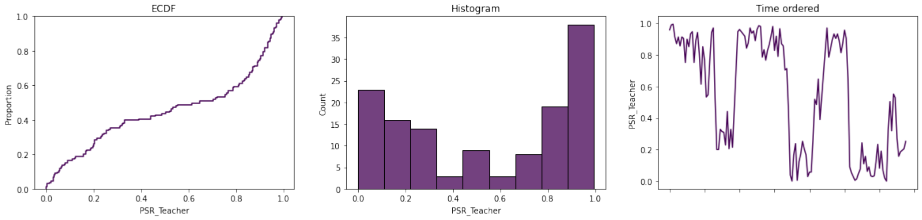


Fig. 5: Analysis of teacher participation rate for the whole course

a classroom profile with information about the student participation, degree of interactivity and teaching styles. The preliminary results show that the data visualization is found very useful by the teachers and reveals meaningful insights about the different classroom activities.

As a statement of direction, we would like to build a bigger dataset with additional audio recordings. Moreover, we are working on automatizing our audio pipeline to generate data reports for each teacher, with customized information for each teaching activity. We are also applying clustering techniques based on machine learning to identify groups of teachers with similar discourse behaviors in the classroom. Finally, for the reproducibility of our scientific research results, another step would be the distribution of our dataset and source code in open repositories.

References

1. Bhattacharya, I., Zhang, T., Ji, H., Foley, M., Ku, C., Riedl, C., . . . Welles, B. F. A multimodal sensor-enabled room for unobtrusive group meeting analysis. *ICMI 2018 - Proceedings of the 2018 International Conference on Multimodal Interaction*, pp. 347–355 (2018)
2. Donnelly, P., Blanchard, N., Samei, B., Olney, A. M., Sun, X., Ward, B., . . . D'Mello, S. K. Automatic teacher modeling from live classroom audio. *UMAP 2016 - Proceedings of the 2016 Conference on User Modeling Adaptation and Personalization*, pp. 45–53 (2016)
3. James, A., Chua, Y. H. V., Maszczyk, T., Núñez, A. M., Bull, R., Lee, K., Dauwels, J. Automated classification of classroom climate by audio analysis. *Lecture Notes in Electrical Engineering*, 579, 41–49. (2019)
4. Lai, C., Carletta, J., Renals, S. Modelling Participant Affect in Meetings with Turn-Taking Features. *Workshop on Affective Social Speech Signals.* (2013)
5. Li, H., Kang, Y., Ding, W., Yang, S., Yang, S., Huang, G. Y., Liu, Z. Multimodal Learning for Classroom Activity Detection. *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 9234–9238 (2020)
6. Logan, B. Mel frequency cepstral coefficients for music modeling. *In Proceedings of International Symposium on Music Information Retrieval* (2000)
7. Nguyen, T. D., Cannata, M., Miller, J. Understanding student behavioral engagement: Importance of student interaction with peers and teachers. *The Journal of Educational Research*, 111(2), pp. 163-174 (2018)
8. Owens, M. T., Seidel, S. B., Wong, M., Bejines, T. E., Lietz, S., Perez, J. R., . . . Tanner, K. D. Classroom sound can be used to classify teaching practices in college science courses. *Proceedings of the National Academy of Sciences of the United States of America*, 114(12), pp. 3085–3090 (2017)
9. Park, T. J., Kanda, N., Dimitriadis, D., Han, K. J., Watanabe, S., Narayanan, S. A Review of Speaker Diarization: Recent Advances with Deep Learning. *Computer Speech Language*, vol. 72, p. 101317 (2022)
10. Ramakrishnan, A., Zylich, B., Ottmar, E., Locasale-Crouch, J., Whitehill, J. Toward Automated Classroom Observation: Multimodal Machine Learning to Estimate CLASS Positive Climate and Negative Climate. *IEEE Transactions on Affective Computing* (2021)
11. Rymes, B. *Classroom discourse analysis: A tool for critical reflection.* Routledge (2015)
12. Schlotterbeck, D., Uribe, P., Araya, R., Jimenez, A., Caballero, D. What classroom audio tells about teaching: a cost-effective approach for detection of teaching practices using spectral audio features. *In LAK21: 11th International Learning Analytics and Knowledge Conference*, pp. 132-140 (2021)
13. Su, H., Dzodzo, B., Wu, X., Liu, X., Meng, H. Unsupervised methods for audio classification from lecture discussion recordings. *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, pp. 3347–3351 (2019)
14. Wang, Z., Pan, X., Miller, K. F., Cortina, K. S. Automatic classification of activities in classroom discourse. *Computers Education*, 78, 115-123 (2014)