2023 IEEE International Conference on Teaching, Assessment and Learning for Engineering (TALE) | 978-1-6654-5331-8/23/$31.00 ©2023 IEEE | DOI: 10.1109/TALE56641.2023.10398309

# AI-driven Teacher Analytics: Informative Insights on Classroom Activities

Oscar Canovas
*Dept. Computer Engineering*
*University of Murcia*
Murcia, Spain
ocanovas@um.es

Felix J. Garcia-Clemente
*Dept. Computer Engineering*
*University of Murcia*
Murcia, Spain
fgarcia@um.es

Federico Pardo
*Dept. Computer Engineering*
*University of Murcia*
Murcia, Spain
federico.pardog@um.es

***Abstract*—Teachers require reliable feedback to refine their skills and to analyze their methods in the classroom. However, it can be challenging for teachers to be aware of the characterization and distribution of classroom activities while simultaneously delivering lectures. Traditional methods of achieving expertise through deliberate practice under the guidance of a human coach are not suitable, due to the long turnover time required for training coaches, observing classrooms, and coding activities. To address these challenges, automated approaches using artificial intelligence (AI) techniques to analyze audio recordings have been proposed to infer the climate in classrooms, model teacher discourse, and classify teaching activities. While these approaches have shown promising results inferring activities, there is a need for additional proposals that support tools and analytics enabling teachers to reflect on their practice and track their progress. In this paper, we present a novel framework that leverages deep learning for speaker diarization and machine learning algorithms for the classification of teaching practices and the analysis of different teaching styles. Our approach utilizes several non-verbal discursive features to provide informative insights. Specifically, we have defined 12 different features that are employed to classify up to three distinct practices: lecture, group discussion, and the use of audience response systems. We show that those features are also informative to analyze the behavior of the teachers for each teaching practice.**

***Index Terms*—Audio analysis, teaching practices, speaker diarization, teaching analytics, deep learning, machine learning**

## I. Introduction

Timely feedback is crucial for teachers to become effective educators. They need to engage in continuous learning, reflect on their practice, and adapt their teaching methods accordingly. Deliberate practice, guided by coaches and accompanied by constructive feedback, is often time-consuming and impractical [1].

Realizing the importance of feedback in teacher learning, it is crucial to provide teachers with a clear and deep understanding of their own performance and progress. We propose an approach that aims to address the lack of immediate and objective feedback. Our approach focuses on providing teachers with timely, accurate, and objective feedback to support their professional growth and enhance their teaching practice. Existing findings suggest that when teachers are provided with automated feedback on the ratio of teacher to student talk, there is a significant increase in the amount of student talk [2]. This indicates that even basic information about teachers' own classroom discourse patterns can lead to positive changes in the desired direction.

The proposal described in this paper is mainly based on features derived from the diarization process [3]. Diarization allows for the segmentation of audio recordings into speaker-specific segments, enabling further analysis of non-verbal cues such as speaker turn duration or speaker overlap. By extracting 12 non-verbal features, we gain valuable insights into the dynamics of teacher discourse. Non-verbal cues provide a rich source of information about the interaction and dynamics within the classroom. Speaker turn duration, for example, can indicate the distribution of speaking turns between the teacher and students, and can be used to derive more accurate features measuring the balance of participation. Speaker overlap can reveal instances of interruption or overlapping speech, which can reflect the level of engagement or conversational dynamics.

One of our aims is to explore the potential of using the feature data to create and validate a model that replicates human coders' identification of classroom activities. We specifically investigate the possibility of developing a classification model that can differentiate between three types of classroom activities: lecturing, student group work, and the use of an audience response system.

Given the existence of various teaching styles across different classroom activities, we will conduct an analysis to examine whether our extracted features can effectively identify differences among individual teachers or groups of teachers who exhibit similar teaching approaches. By investigating the distinctive patterns and behaviors captured by our features, we aim to gain insights into the variations in instructional methods employed by educators within each specific classroom activity.

Consequently, in this paper, we pose the following research questions:

- RQ1: Can the features derived from the diarization process effectively identify different teaching practices?
- RQ2: Are the features derived from the diarization process suitable for differentiating between teaching styles of different teachers, within a particular teaching practice?

As we will show, our main contribution is not only providing teachers with an AI-driven mechanism to identify the classroom activities they are conducting, but also enabling them to analyze their discourse patterns and compare those

patterns with data from other teachers. This information provides valuable insights for teachers to reflect on their practices in a timely manner.

## II. Related Work

In recent years, a number of studies have been conducted to analyze the classroom climate [4] and discourse in various contexts. Our primary focus is the automated analysis of teacher discourse, particularly through the utilization of recorded audio. While the practice of recording classroom sessions for teacher assessment is not new, the transition to automatically analyze these recordings has only recently gained momentum [5]. Most of the recent proposals apply machine or deep learning techniques to examine different teaching practices and styles. The analysis of such recordings can involve the extraction of non-verbal features or the application of natural language processing techniques. The proposal described in this paper is mainly based on features derived from the diarization process, that is, labeling the speakers and their respective speaking instances. This task can be approached using various methods, ranging from traditional approaches to advanced neural networks [3].

In relation to classroom discourse, multiple research teams have dedicated their efforts to the development and validation of automated systems aimed at discerning fundamental discourse structures, such as lectures and group work. For instance, Donnelly et al. [6] trained supervised machine learning models to classify instructional segments, achieving F1 scores ranging from 0.64 to 0.78. Furthermore, [7] demonstrated the feasibility of employing automatic speech recognition and classification models to automatically segment classroom speech and identify instances where teachers' utterances contained questions. Other works based on automatic speech recognition have focused on segmenting teacher and student classroom speech [8] and leveraging low-level acoustic features. Most of the proposals pay particular attention to the role of the teacher [9] in order to classify active learning tasks [10]–[12].

However, it is worth noting that these works predominantly focus on achieving high classification accuracy, often overlooking the provision of discourse features that can serve as descriptive and informative data [2], [11]. Analyzing different teaching styles and offering valuable recommendations to teachers requires not only accurate classification but also a comprehensive understanding of the underlying discourse dynamics. Informative data are crucial for capturing the nuances of teaching practices and providing meaningful insights. We believe the research presented in this paper is the first to use informative features to perform not only classification of teaching practices but also an analysis of teaching styles within each method.

As we will introduce in IV-C, we adapt some of the discourse features from previous works such as [13], [14] which were originally designed for group meeting analysis. We find that certain turn-taking features, participation rates, and silence-related features presented in these works are also applicable to the analysis of teaching practices.

Currently, there are commercial products available, such as the start-up TeacherFX[1], which has primarily focused on quantifying the ratio of teacher-to-student talk and providing graphs related to participation and interaction. However, their approach is limited in terms of the number of features analyzed, preventing a detailed and comprehensive analysis that we aim to achieve with our research.

In addition to the non-verbal approach, traditional methods of automated teacher discourse analysis have relied on automatic speech recognition (ASR) transcripts [15]. This widely adopted method involves extracting high-level linguistic features spanning word, sentence, and discourse levels. These features are then used as inputs to supervised classifiers, enabling the detection of desired discourse features in a generalizable manner. The trained classifiers function as the initial points of reference for subsequent fine-tuning [16], wherein they are augmented with a task-specific output layer and fine-tuned by updating the parameters using limited amounts of domain-specific data. Notably, in some studies [17], deep learning techniques have been employed, for example, to identify particular dialogic strategies within mathematics classrooms. While we are now working on a hybrid system for classroom analysis based on non-verbal features and NLP (Natural Language Processing) techniques, this latter approach is out of the scope of this paper.

## III. Method

### A. Datasets

In this study, we conducted an analysis of audio recordings obtained from two courses, namely Computer Networks and Computing Foundations, which are part of a bachelor's degree program in Computer Science. To ensure a comprehensive dataset, we engaged 1 female and 3 male teachers to record their respective classes, resulting in a total of 38 audio files, with each file corresponding to an individual class taught by one of the teachers. Teachers in our sample averaged 20.5 years of experience. The cumulative duration of these audio files is approximately 24 hours, with each file ranging from 1 to 2 hours in duration. All the data was collected with the approval of the teachers, students and the Institutional Review Board.

Furthermore, each audio session included supplementary metadata. This metadata encompassed information such as the specific course to which the audio file pertained, the date on which the recording took place, the duration of the audio file, the number of students present during the recording, the recording device utilized, the teaching methods employed during the class session, and the identity of the instructor delivering the lecture.

There are 4 labels for the segments of the recordings:

- **Lecture**. This represents a traditional class where the teacher explains the course material, and the students

[1]https://teacherfx.com

listen to the teacher, take notes, and ask questions occasionally.

- **Wooclap**. This label is used when an Audience Response System [18]is employed in the classroom to pose questions to the students and engage with them through the use of mobile devices. Wooclap[2] is the platform utilized in all instances.
- **Group Work**. This represents a session where the students work in groups to solve problems.
- **Other**. This label is used when the audio segment does not align with any preceding description, such as instances when the teacher is getting ready to start or during a break in the class. The segments with this label are omitted for this work.

### B. Human coding

As mentioned earlier, we have stored metadata concerning the teaching methods employed in each class; however, this information lacks time marks. Therefore, we required a timestamp system to associate the teaching practices with the corresponding audio segments. The labeling method adopted consists of an audio timestamp indicating the start of the label, another timestamp denoting the end of the labeled segment, and the corresponding teaching method. This manual process was performed by a single independent coder using the ELAN software[3]. The resulting data serves as the ground truth for training the various machine learning models that were tested for automated classification of teaching practices. According to the coder, the dataset is composed by 13 hours of lecture, 3 hours of Wooclap, and 8 hours of work group.

### C. Experimental setup

During an in-person training session, teachers were provided instructions on how to record a class session. As part of the process, they conducted an initial recording in their respective classrooms for test purposes and to ensure optimal placement of the digital recorder (TASCAM DR-07X). Each teacher was required to record a minimum of ten sessions encompassing various teaching practices. From our dataset, we specifically chose recordings that exhibited high audio quality for further analysis.

These classes were held in traditional lecture classrooms with a seating capacity of 60-90 students. The classrooms are equipped with tables and chairs, facing the teaching area which includes a desk with fully embedded technology. To ensure a seamless recording setup, a handheld digital recorder was discreetly placed on the teacher's desk, positioned at least 1.5 meters away from both the teacher and the students in the front row.

## IV. Proposed Approach

Our approach consists of a five-stage pipeline, as shown in Figure 1. Each stage provides input to the subsequent stage to obtain more valuable information. The Data Visualization stage specifically offers visual information to teachers for actionable teaching analytics.

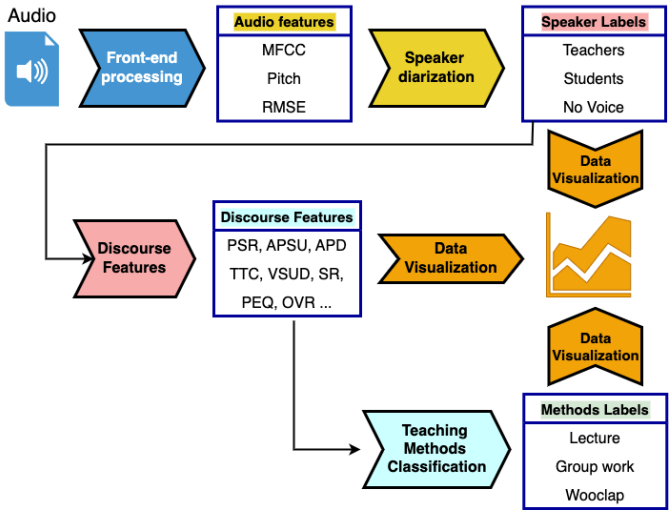



Fig. 1: Pipeline of processing stages and data elements

### A. Front-end processing

The initial stage of audio processing involves extracting low-level features from the audio recordings. Some important features include Mel Frequency Cepstral Coefficients (MFCCs) [19], pitch, and average energy for each voiced frame. These features can be obtained using various tools such as Torchaudio[4] or Librosa[5], among others. These features are widely employed in scientific literature as inputs for speaker diarization techniques.

### B. Speaker diarization

Speaker Diarization is the process of dividing audio recordings into segments and assigning speaker labels to determine "Who Speaks When". A diarization system comprises a Voice Activity Detection (VAD) model that identifies time intervals in the audio where speech is present while ignoring background noise. It also includes a Speaker Embeddings model that extracts unique speaker audio features from the speech segments identified by the VAD. In our approach, we are able to differentiate the teacher's speeches; however, we do not distinguish between individual students, labeling all other speakers as "Students". Achieving precise speaker diarization can be accomplished using tools like Pyannote-audio[6], which is the option that we used in our experiments.

### C. Feature calculation

Considering some previous works [13], [14], we have defined several non-verbal features extracted from the diarization that are useful for the characterization of discourses. Specifically, we define four discourse features per role (teacher and students):

[2]https://wooclap.com
[3]https://archive.mpi.nl/tla/elan
[4]https://pytorch.org/audio/stable/transforms.html
[5]https://librosa.org/doc/main/feature.html
[6]https://github.com/pyannote/pyannote-audio


Fig. 2: Timelines of different teaching practices: (top) lecture; (middle) work group; (bottom) Wooclap

- **Participant Speaking Ratio (PSR)**. The ratio of participation of each role during the recording segment.
- **Participant Speaking Utterances (PSU)**. The number of utterances in the current recording segment.
- **Participant Speaking Utterances Ratio (PSUR)**. The ratio of utterances of each role during the recording segment.
- **Average Participant Speaking Utterance Duration (APSUD)**. The average duration of the utterances of each role.

Additionally we have also defined eight global discourse features:

- **Average Lapse Duration (ALD)**. The average duration of the period of silence.
- **Silence Ratio (SR)**. The ratio of the period of silence during the recording segment.
- **Average Pause Duration (APD)**. The duration of the average silence interval between utterances by the same participant.
- **Participation Equality (PEQ)**. An indicator that assesses the balance of participation among different roles. It is calculated following the methodology outlined in [14]. Values close to 1 indicate an equal distribution of participation.
- **Turn Taking Count (TTC)**. The number of turn changes occurred in the dialogue between students and the teacher.
- **Very Short Utterances Ratio (VSUR)**. The ratio of very short speech utterances (less than 2 seconds) over the total.
- **Overlapping Rate (OVR)**. The rate of overlapping utterances of different participants.
- **Overlapping Utterances Rate (OVUR)**. The ratio of utterances that are overlapped.

### *D. Teaching methods classification*

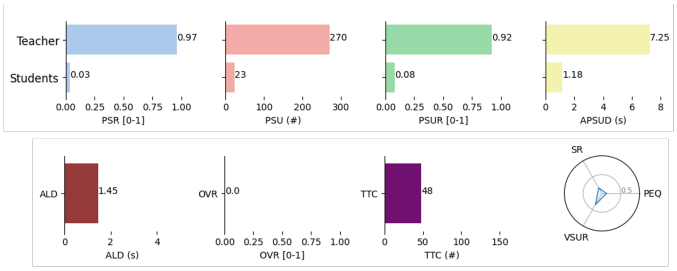
We performed a multiclass classification task involving three classes: Lecture, Group Work, and Wooclap. Our approach employed supervised machine learning algorithms to predict the appropriate label for a given input recording. We explored a range of models, including Support Vector Machines (SVM), k-Nearest Neighbor (kNN), Random Forest, Naive Bayes, Logistic Regression, and Gradient Boosting, to identify the most effective algorithm for our task.

As a teacher can use multiple teaching methods within a single class, our classification stage aims to identify the start and end points of each method. To achieve this, the classification process involves analyzing small segments of the recordings. We have examined overlapping window sizes ranging from 60 seconds to 300 seconds. Our preliminary results show that the best trade-off between granularity and accuracy is obtained using windows of 120 or 180 seconds. More details about the experimental results are shown in Section V.

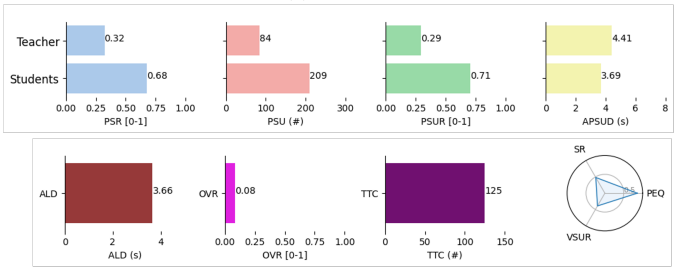
### *E. Data visualization*

Effective data visualization is crucial in providing teachers with valuable insights about their teaching activity and equipping them with actionable analytics to enhance their teaching practices.

Our approach involves presenting tailored data visualizations for each stage of the pipeline, as we already presented in [removed for anonymity]. Specifically, we propose employing timeline graphs to depict the output of the diarization process. As Figure 2 shows, timeline visualization offers a clear depiction of the chronological sequence of participants, helping teachers in understanding temporal relationships. Moreover, timelines are also useful to analyze the different interaction




(a) Lecture




(b) Wooclap session

Fig. 3: Example of feature visualization for two recordings of different teaching practices.

patterns that each teaching practice exhibits. Here, T02 refers to the teacher and SPEAKER_0X to different students.

To represent discourse features, we propose utilizing bar charts and radar plots to facilitate comparisons between multiple features. As depicted in Figure 3, certain features exhibit divergent values based on the employed teaching method, which provides an encouraging foundation for investigating the potential of these features in automating the classification of teaching practices.

Finally, after labeling the audio segments based on the teaching practice being employed, we can display information regarding the various features within the context of each teaching method, for example using box plots. Furthermore, as we show in Section V-B, we can compare the behavior of different teachers for each feature, aiming to identify distinct teaching styles and provide valuable insights to teachers about their abilities.

## V. Results and Discussion

### A. (RQ1) Identification of teaching practices

We have established a set of 12 distinct non-verbal features derived from the diarization process, and our objective is to assess their suitability for automating the classification of teaching practices. To initiate this investigation, we employ unsupervised techniques to explore the data. In particular, we employ dimensionality reduction techniques such as t-SNE (t-distributed Stochastic Neighbor Embedding) to reduce the data's dimensionality and facilitate visualization.

By applying t-SNE to the data (with three PCA components and exaggeration=1), we obtain the visualization shown in Figure 4. As we observe, it confirms the potential separability

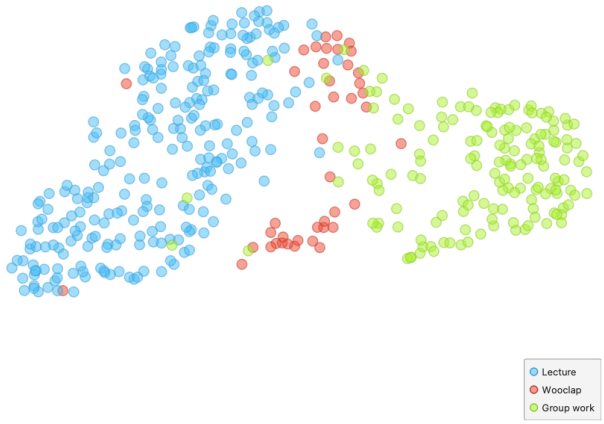



Fig. 4: t-SNE visualization of three distinct clusters representing teaching practices.

of clusters within the data. The distribution of samples demonstrates distinct groupings that indicate promising prospects for classifying different teaching practices automatically, with challenging conditions for the "Wooclap" label.

In our study, we thoroughly investigated several machine learning techniques to determine their effectiveness in classifying the features derived from audio samples of 180 seconds.

To ensure comprehensive evaluation, we employed data from all the teachers participating in the experiment and all the teaching methods. The classification performance of each technique was assessed by means of a cross-validation process using 10 folds. This approach involved splitting the dataset into ten equal parts, performing the classification on nine of the folds while using the remaining fold as the validation set. This process was repeated ten times, with each fold serving as the validation set once.

To provide reliable and representative results, we averaged the performance metrics obtained from the cross-validation process over the three distinct classes being classified. The resulting values are presented in Table I.

TABLE I: AUC, F1, Precision and Recall for different models over the cross validation using optimal parameters.

| Model | AUC | F1 | Precision | Recall |
|---|---|---|---|---|
| kNN | 0.973 | 0.9288 | 0.9304 | 0.9279 |
| SVM | 0.9824 | 0.9373 | 0.9386 | 0.9364 |
| Random Forest | 0.9832 | 0.8945 | 0.8973 | 0.9067 |
| Naive Bayes | 0.978 | 0.8743 | 0.9153 | 0.8559 |
| Logistic Regression | 0.9873 | 0.9342 | 0.9342 | 0.9343 |
| Gradient Boosting | 0.9801 | 0.9252 | 0.9246 | 0.9258 |

Among the evaluated models, the SVM model exhibited the best performance, achieving an F1 score close to 94%. This indicates the robustness and effectiveness of the SVM model in accurately classifying the teaching practices based on

the extracted features. The high F1 score represents a balance between precision and recall, highlighting the model's ability to correctly identify relevant instances while minimizing false positives and false negatives.

In addition to the performance metrics, the confusion matrix (Table II) provides insights into the misclassification patterns of the SVN model. We observe which specific cases are being misclassified, in this case mainly those related to the "Wooclap" method. The "Wooclap" category achieved an accuracy of 75.6% in our classification analysis. Upon examining the misclassified instances, we observed that the errors were almost evenly distributed between the other two labels. This observation aligns with the findings from the t-SNE representation, indicating that the audio patterns of Wooclap sessions sometimes exhibit similarities to either lectures or group work, contributing to the misclassifications.

TABLE II: Confusion matrix for the SVM model

| Actual / Predicted | Lecture | Wooclap | Group Work | Σ |
|---|---|---|---|---|
| **Lecture** | 97.3% | 2.3% | 0.4% | 260 |
| **Wooclap** | 11.1% | 75.6% | 13.3% | 45 |
| **Group Work** | 1.8% | 5.4% | 92.8% | 167 |
| **Σ** | 261 | 49 | 162 | 472 |

### *B. (RQ2) Differentiated teaching styles*

Once we have confirmed that using the non-verbal features, we are able to classify the teaching practices employed in each audio segment with a notably good F1 score, it is time to analyze their informative counterpart. As we stated in our RQ2, we are going to examine whether those features are suitable for differentiating teaching styles within a specific teaching practice. Furthermore, we want to analyze if those differences are informative for the teachers and can be used as actionable analytics.

Since there are 12 different features, providing a detailed analysis of all of them would require excessive space in this paper. Therefore, we have focused our attention on one global feature related to silences, namely "ALD" (average lapse duration), and two features primarily associated with how the teachers conduct a session: "APSUD Teacher", indicating the average duration of teacher utterances within the segment, and "PSR Teacher", which represents the ratio of teacher participation during the audio segment. Those features will be analyzed for two different teaching practices: lecture and Wooclap. These methods were selected because they involve a more predominant role of the teacher compared to group work sessions.

Firstly, we will examine the statistical distribution of these features. Figure 5 displays the boxplots generated from the dataset, representing all samples for each teacher and the selected teaching practices.

On the one hand, considering lectures, the box plots in Figure 5b depict the average duration of utterances (APSUD) during the different audio segments for the four teachers. Upon careful examination, it becomes clear that one of the teachers exhibits a significantly higher distribution compared to the remaining teachers. This particular teacher consistently presents longer average utterance durations across the analyzed segments, suggesting a distinct speaking pattern and potentially indicating a unique teaching style. The noticeable deviation in the box plot emphasizes the importance of considering individual teacher features and their impact on the overall dynamics of the classroom environment. A similar behavior is also observed for the participant speaking ratio (PSR) for the same teacher.

To further validate the observed differences, a statistical analysis is required. The distributions are confirmed to be non-normal. Therefore, the non-parametric Kruskal-Wallis test was employed. Table III provides the details of this analysis for lectures.

TABLE III: Kruskal-Wallis test results for three features among four teachers for the lecture sessions

| Variable | $\chi^2$ | gl | p | $\epsilon^2$ |
|---|---|---|---|---|
| ALD | 81.5 | 3 | <.001 | 0.315 |
| APSUD Teacher | 88.2 | 3 | <.001 | 0.340 |
| PSR Teacher | 119 | 3 | <.001 | 0.461 |

The test results indicate significant differences among the teachers for all three variables: "APSUD Teacher", "PSR Teacher", and "ALD". The effect sizes ($\epsilon^2$) suggest moderate to substantial impacts of the teachers on the variables.

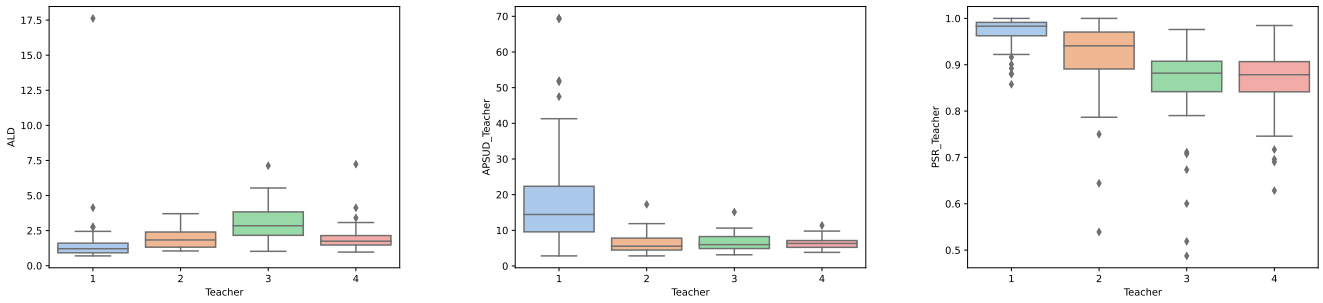
On the other hand, considering Wooclap sessions, the box plots displayed in Figure 5(d-f) reveal an interesting pattern where two distinct groups of teachers with similar distributions can be observed. The first group, consisting of Teachers 1 and 4, shows comparable average lapse durations and participant speaking ratios across different audio segments. Similarly, the second group, comprising Teachers 2 and 3, also exhibits similar distribution patterns. These findings suggest two different teaching styles. The statistical analysis is therefore required to confirm the significance of these observed groupings and determine if they are statistically meaningful.

TABLE IV: Kruskal-Wallis test results for three features among four teachers for the Wooclap sessions

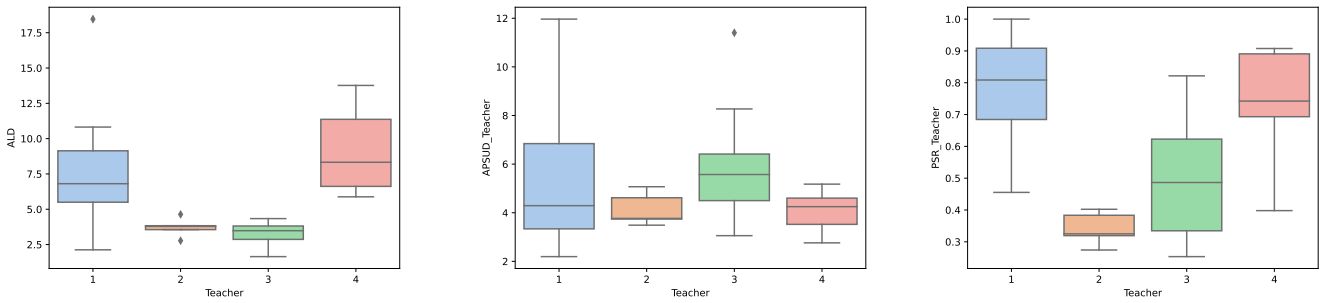
| Variable | $\chi^2$ | gl | p | $\epsilon^2$ |
|---|---|---|---|---|
| ALD | 27.9 | 3 | <.001 | 0.635 |
| APSUD Teacher | 6.53 | 3 | .088 | 0.149 |
| PSR Teacher | 24 | 3 | <.001 | 0.545 |

The Kruskal-Wallis test results suggest that there are significant differences among the teachers for the variables "PSR Teacher" and "ALD" in the context of Wooclap sessions. However, there is no significant difference for the "APSUD Teacher" variable. The effect sizes ($\epsilon^2$) indicate that the teachers have a moderate to large impact on the variables.

The pairwise comparisons using the Dwass-Steel-Critchlow-Fligner test were conducted to confirm the presence of two distinct groupings among the teachers. The results, as shown in Table V, reveal that for the "PSR Teacher" feature there are no differences between T01 and T04, and T02 and T03.




(a) ALD (Lecture) (b) APSUD Teacher (Lecture) (c) PSR Teacher (Lecture)




(d) ALD (Wooclap) (e) APSUD Teacher (Wooclap) (f) PSR Teacher (Wooclap)

Fig. 5: Distributions of the features ALD, APSUD Teacher and PSR Teacher for the participating teachers for Lecture sessions (top) and Wooclap sessions (bottom).

TABLE V: Pairwise Comparisons of PSR Teacher using Dwass-Steel-Critchlow-Fligner test

| Comparison PSR Teacher | W | p |
|---|---|---|
| Teacher T01 vs Teacher T02 | -4.475 | 0.008 |
| Teacher T01 vs Teacher T03 | -4.762 | 0.004 |
| Teacher T01 vs Teacher T04 | -0.616 | 0.972 |
| Teacher T02 vs Teacher T03 | 1.913 | 0.529 |
| Teacher T02 vs Teacher T04 | 4.391 | 0.010 |
| Teacher T03 vs Teacher T04 | 5.179 | 0.001 |

### C. Discussion

Both research questions have been answered affirmatively in our study. Firstly, we successfully demonstrated that the features derived from the diarization process effectively identify different teaching practices. Additionally, we have proven that these same features are suitable for distinguishing between the teaching styles of different teachers within a specific teaching practice. Furthermore, our proposal offers teachers valuable data visualization for each stage of the analysis.

We believe that data visualization based on non-verbal discursive features provides informative insights into classroom activities. However, teachers may require a point of reference or guidance to improve their teaching activities. This point of reference could be an ideal teaching style defined by experts or even an AI model generated using a relevant dataset that combines teaching styles, learning outcomes, and academic results. In cases where obtaining an ideal teaching style is unfeasible, this point of reference could also be derived by calculating the mean values of the feature set for an academic community. Both novel and experienced teachers concerned about their teaching style need a reliable point of reference to guide their improvement actions. Our proposed data visualization mechanism serves as a valuable tool to address this need and facilitates decision-making.

For example, the box plot of the feature APSUD for Lecture sessions in Figure 5b reveals that the value of this feature for teacher 1 is significantly higher than the values for the other teachers. This observation suggests that an automated recommender system could suggest that teacher 1 should consider shortening their utterances. Furthermore, these recommendations could be provided as guidance collectively. For instance, assuming that teacher 4 represents an ideal teaching style, the recommender system could propose, based on the data visualization of the box plots for Lecture sessions in Figure 5, that teacher 1 should increase the duration of silent periods and encourage student participation in class, in addition to shortening their utterances.

Furthermore, the statistical analysis conducted on our reduced dataset has confirmed the significance of the observed groupings of teaching styles. These groupings hold valuable implications for various purposes. Firstly, they can be utilized to define customized AI models tailored to the unique profile of each teacher. By incorporating the specific characteristics and preferences associated with their teaching style, AI models can provide personalized support and guidance. Secondly, these identified teaching style groupings can serve as a foundation for designing further investigations. By categorizing the teachers according to their teaching styles, it allows for

conducting more targeted and focused investigations, aiming to explore the specific instructional approaches, strategies, and outcomes associated with each teaching style in greater detail.

## VI. Conclusions and Future Work

In conclusion, this paper presents an approach that leverages the diarization process and extracts 12 non-verbal features from audio recordings to provide teachers with valuable insights into their teaching practices. The findings highlight the potential of these features in differentiating between various teaching practices and identifying distinct teaching styles among individual teachers. The proposed AI-driven mechanism empowers teachers to analyze their own discourse patterns, compare them with peers, and gain timely feedback to enhance their professional growth.

While our proposed system holds the potential for evaluating teachers and promoting adherence to specific teaching practices, it is important to note that our current research does not incorporate standardized measures for assessing teaching quality or enforcing particular practices. Instead, we provide teachers with informative features to facilitate reflection on their discourse patterns. Further analysis, such as correlating student performance or positive perceptions with our computed features, could enable AI-driven recommendations, enhancing the potential impact of our system.

Another area for improvement is related to the accuracy of the classification models. Teacher discourse is inherently sequential in that the current utterance is strongly influenced by the previous set of utterances. Thus, one method of improvement is to use deep sequence learning models, able to learn long temporal dependencies in the data. We will also consider representational learning techniques by using word embeddings, obtained by encoding words as vectors that capture the similarity with other words.

As a final statement of direction, we acknowledge the importance of addressing the limited diversity in our sample. To overcome this limitation, we have planned a new phase of our research that will involve collecting additional data from a more diverse group of teachers. To broaden the scope of our data collection, we will also include widely available recordings from streaming platforms. This approach will enable us to capture a more representative range of teaching practices and enhance the generalizability of our findings.

## Acknowledgments

The authors are grateful for the rater and teachers who participated in this study and helped to collect the data. This work has been funded under grant TED2021-129300B-I00, by MCIN/AEI/10.13039/501100011033, NextGenerationEU/PRTR, UE, and grant PID2021-122466OB-I00, by MCIN/AEI/10.13039/501100011033/FEDER, UE.

## References

[1] J. Archer, S. Cantrell, S. L. Holtzman, J. N. Joe, C. M. Tocci, and J. Wood, *Better feedback for better teaching: A practical guide to improving classroom observations*. John Wiley & Sons, 2016.

[2] Z. Wang, X. Pan, K. F. Miller, and K. S. Cortina, "Automatic classification of activities in classroom discourse," *Computers & Education*, vol. 78, pp. 115–123, 2014.

[3] T. J. Park, N. Kanda, D. Dimitriadis, K. J. Han, S. Watanabe, and S. Narayanan, "A review of speaker diarization: Recent advances with deep learning," *Computer Speech and Language*, vol. 72, 2022. [Online]. Available: http://arxiv.org/abs/2101.09624

[4] A. James, Y. H. V. Chua, T. Maszczyk, A. M. Núñez, R. Bull, K. Lee, and J. Dauwels, "Automated classification of classroom climate by audio analysis," in *9th International Workshop on Spoken Dialogue System Technology*. Springer, 2019, pp. 41–49.

[5] M. E. Dale, A. J. Godley, S. A. Capello, P. J. Donnelly, S. K. D'Mello, and S. P. Kelly, "Toward the automated analysis of teacher talk in secondary ela classrooms," *Teaching and Teacher Education*, vol. 110, p. 103584, 2022.

[6] P. J. Donnelly, N. Blanchard, B. Samei, A. M. Olney, X. Sun, B. Ward, S. Kelly, M. Nystran, and S. K. D'Mello, "Automatic teacher modeling from live classroom audio," in *Proceedings of the 2016 conference on user modeling adaptation and personalization*, 2016, pp. 45–53.

[7] P. J. Donnelly, N. Blanchard, A. M. Olney, S. Kelly, M. Nystrand, and S. K. D'Mello, "Words matter: automatic detection of teacher questions in live classroom discourse using linguistics, acoustics, and context," in *Proceedings of the Seventh International Learning Analytics & Knowledge Conference*, 2017, pp. 218–227.

[8] S. K. D'Mello, A. M. Olney, N. Blanchard, B. Samei, X. Sun, B. Ward, and S. Kelly, "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, pp. 557–566.

[9] D. Schlotterbeck, P. Uribe, R. Araya, A. Jimenez, and D. Caballero, "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*, 2021, pp. 132–140.

[10] H. Li, Y. Kang, W. Ding, S. Yang, S. Yang, G. Y. Huang, and Z. Liu, "Multimodal learning for classroom activity detection," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 9234–9238.

[11] M. T. Owens *et al.*, "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*, vol. 114, no. 12, pp. 3085–3090, 2017.

[12] H. Su, B. Dzodzo, X. Wu, X. Liu, and H. Meng, "Unsupervised methods for audio classification from lecture discussion recordings." in *INTERSPEECH*, 2019, pp. 3347–3351.

[13] I. Bhattacharya, M. Foley, N. Zhang, T. Zhang, C. Ku, C. Mine, H. Ji, C. Riedl, B. F. Welles, and R. J. Radke, "A multimodal-sensor-enabled room for unobtrusive group meeting analysis," in *Proceedings of the 20th ACM International Conference on Multimodal Interaction*, 2018, pp. 347–355.

[14] C. Lai, J. Carletta, and S. Renals, "Modelling participant affect in meetings with turn-taking features," in *Proc. Workshop of Affective Social Speech Signals*, 2013.

[15] T. Nazaretsky, J. N. Mikeska, and B. Beigman Klebanov, "Empowering teacher learning with ai: Automated evaluation of teacher attention to student ideas during argumentation-focused discussion," in *LAK23: 13th International Learning Analytics and Knowledge Conference*, 2023, pp. 122–132.

[16] A. Suresh, J. Jacobs, M. Perkoff, J. H. Martin, and T. Sumner, "Fine-tuning transformers with additional context to classify discursive moves in mathematics classrooms," in *Proceedings of the 17th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2022)*, 2022, pp. 71–81.

[17] A. Suresh, T. Sumner, J. Jacobs, B. Foland, and W. Ward, "Automating analysis and feedback to improve mathematics teachers' classroom discourse," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 33, no. 01, 2019, pp. 9721–9728.

[18] J. I. Castillo-Manzano, M. Castro-Nuño, L. López-Valpuesta, M. T. Sanz-Díaz, and R. Yñiguez, "Measuring the effect of ars on academic performance: A global meta-analysis," *Computers & Education*, vol. 96, pp. 109–121, 2016.

[19] B. Logan *et al.*, "Mel frequency cepstral coefficients for music modeling." in *Ismir*, vol. 270, no. 1. Plymouth, MA, 2000, p. 11.