Teaching practices analysis through audio signal processing

Braulio Ríos[1], Emilio Martínez[1], Diego Silvera[1],
Pablo Cancela[1], and Germán Capdehourat[1,2]

[1] Instituto de Ingeniería Eléctrica, Facultad de Ingeniería,
Universidad de la República, Uruguay
{braulio.rios,emartinez,dsilveracoeff,pcancela}@fing.edu.uy
[2] Ceibal, Uruguay
gcapdehourat@ceibal.edu.uy

Abstract. Remote teaching has been used successfully with the evolution of videoconference solutions and broadband internet availability. Even several years before the global COVID 19 pandemic, Ceibal used this approach for different educational programs in Uruguay. As in face-to-face lessons, teaching evaluation is a relevant task in this context, which requires many time and human resources for classroom observation. In this work we propose automatic tools for the analysis of teaching practices, taking advantage of the lessons recordings provided by the videoconference system. We show that it is possible to detect with a high level of accuracy, relevant lessons metrics for the analysis, such as the teacher talking time or the language usage in English lessons.

Keywords: teaching analysis · classroom activity detection · diarization · education · audio signal processing

1 Introduction

Classroom observation and teaching evaluation have historically been relevant activities in the field of education [9]. In this context, many efforts have been made to standardize lesson observation [6]. In all existing observation protocols, the effort required to apply them is still very high, mainly due to the time and human resources involved in the task to implement it on a large scale. To tackle this scalability limitation, in this work we propose and validate different automatic classroom observation tools to assist the analysis of remote teaching practices, based on the processing of the lessons recordings.

The increasing deployment of broadband internet access at schools enabled new ways of teaching, such as remote lessons through videoconference solutions. In Uruguay, even several years before the global COVID-19 pandemic, this technology was implemented for different educational programs [1]. Ceibal, an organization that provides technological support to the K-12 education system in Uruguay, for example used this approach to universalize the English lessons as a second language at the primary education level. The main problem was the lack of local English teachers, which was solved in a joint work with the British Council [7], which provided the required teachers that are placed all over the world.

This innovative educational approach with remote teachers for English lessons was very successful. Thus, this methodology was also later extended for Computational Thinking courses. One of the key points of the remote English lessons program, that has been addressed from the very beginning, is the continuous quality monitoring process of the lessons and teachers involved. A group of education technicians, the so-called *quality managers*, attend every year to some lessons from different teachers, following a standardized observation protocol that allows them to review the different activities carried out. After each lesson observation, they write a report to give feedback and exchange ideas with the remote teachers. Their work contributes to the continuous improvement of pedagogical practices, which allows to identify strengths and weaknesses of the academic program and thus plan enhancements for the following year.

The limited number of quality managers, together with the great amount of time that an observation requires, only allows to monitor a reduced number of lessons throughout the year. Therefore, any automation that could be introduced to support their work, would have a great impact in the information available to analyze and improve the educational program. In this context, we propose different tools implemented with state-of-the-art audio processing techniques applied to the lessons recordings, that would be of great help in this regard.

As we show in the following sections, the results obtained validate the utility of the proposed tools for the automation of different relevant analyses of teaching practices observed during a lesson. In the next section, we review some previous work in the area related to classroom analysis. Then, in Section 3 we describe the dataset that was built for this work. Next, Section 4 focuses on classroom activity detection, addressing the problem of detecting whether the teacher or students are talking at each moment of the lesson. Additional tools are presented in Section 5, such as the identification of the spoken language during a lesson and the detection of key phrases, related to the particular content of the educational Unit that should be covered. Finally, the paper ends with Section 6, presenting the main conclusions and new lines of work that could be studied in further research.

2 Related Work

With the advances in machine learning, the automation of classroom activity detection have been addressed in several previous works. One of them is the Decibel Analysis for Research in Teaching (DART) [10], a simple approach based on the power of the audio signal. This method detects the lesson segments with only one voice, with more than one voice speaking simultaneously, and with no voices at all. The authors report an accuracy close to 90% for college classes.

More recent works are mostly based on deep learning techniques [4,13,14]. The typical approach is to first extract more powerful low level features rather than just the audio signal power, such as the cepstral coefficients of the Mel bands (MFCC). These features have proven to be very versatile and provide good results in a wide variety of applications, particularly in speech processing [8]. The next step is to train a neural network, defining suitable labels for the stated problem. Typical tags could be the same as in DART (no voice, single voice or multiple voices) or other higher level labels, such as the ones used in [13] to identify teaching practices (e.g. presenting, guiding or administration tasks), where they report 80% of accuracy for the two most common categories.

All the previous work found is based on supervised learning. To the best of our knowledge, this is the first paper to use diarization techniques to tackle the classroom activity detection problem. Speaker diarization responds to the question of *who spoke when* in an audio signal and deep learning techniques have significantly improved the performance of state-of-the-art algorithms for this purpose [11,15]. As we will see in more detail in Section 4, it is possible to directly apply unsupervised diarization for teaching analysis, thus avoiding the costly data labeling for training. Although the results obtained are worse than for the supervised approach, they could still be useful for various applications.

In addition to speaker analysis, in this paper we also propose to use state-of-the-art techniques for the identification of the spoken language throughout the lesson and the detection of key phrases. Language identification is a long date relevant problem in the audio processing community [5,16] and an important performance boost has been achieved with the recent Transformers based algorithms. Open source models such as Whisper [12], which serve for speech to text transcription, also has quite good performance in the language detection task, as we will show in Section 5. Moreover, we explore the usage of the lesson transcription for key phrases detection, something already studied in previous works [2,17]. This makes it possible to detect the lesson segments where certain specific content of each Unit is addressed, as well as to analyze the kind of feedback that teachers provide to the students.

None of the previous works found refer to the specific use case of remote teaching lessons. There are some particular characteristics of this context that pose specific challenges. On the one hand, the lessons to be analyzed are taught remotely, that is, the students are in a classroom with their local teacher, while the remote teacher guides the lesson through the videoconference system. This fact has an impact in the the quality of the recordings, which are made with the microphones of the videoconference system itself, which are not high-fidelity equipment designed for further signal processing. On the other hand, the lessons dynamics in a primary school environment present considerably differences with respect to the classes taught in secondary or college level. This is a significant difference with other previous works and it also affects the recordings quality for the problem posed, since at the primary level it is more complex to control the behavior of students, and even more so in a lesson which is taught remotely.

3 Dataset description

One of the most important aspects related to any machine learning algorithm development is the appropriate dataset collection and the corresponding data labeling. In this case, the raw data corresponds to the videoconference recordings of the lessons from the English and Computational Thinking courses managed by Ceibal. It is worth noting that both the teachers and the parents or the legal guardians of the students involved, gave their consent for the collection of the data for this research. Although the recordings include the videos of the lessons, only the audios were used for this work.

A carefully specified labeling protocol was defined in order to tag who is talking at each time in the recording of a lesson. Each tag includes its start and end times as well as a label. These labels indicate the presence of the teacher's voice (***teacher***), an individual student voice (***student***) or multiple overlapping student voices (***multiple***) such as during teamwork activities or due to answers in chorus. An example of the labels is illustrated in Figure 1. The labeling protocol also included other higher level tags, such as classroom disorder, particular noises (crawling chairs, knocks on tables, noisy vehicles outside the classroom) and hushing utterances ("*shh*"). The latter were included to have more information about each particular lesson, and being able to relate performance drops observed with audio quality issues in the data.

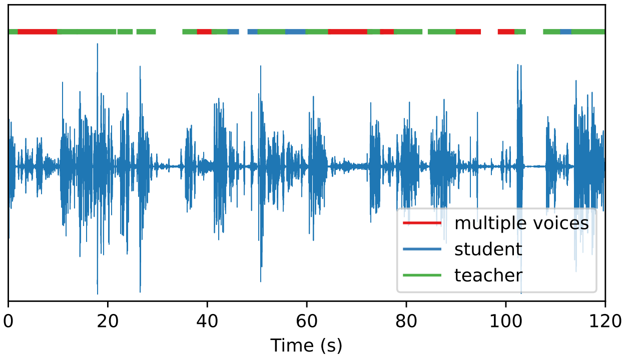


Fig. 1. Example of the labeling result along the raw audio waveform for a lesson.

It should be noted that the data labeling stage was essential for this work, due to the lack of public databases and the different educational context analysed in some previous works, such as university lessons. In order to ensure the quality of the labeling data, and also to minimize ambiguities in the defined protocol, two labeling stages were performed. In the first stage, the three people involved in the labeling process worked with the same six different recordings, which together reached a total of 4.5 hours of teaching time. With the results of the first stage, the differences found between the labels were analyzed and discussed, and finally the labeling criteria were adjusted. Once the final protocol version was defined, the second labeling stage was carried out with all the available lessons recordings, reaching a total of 19 hours of manually labeled audio data, complying with the aforementioned protocol. It is worth to mention that each hour of lesson recording requires approximately one and a half hour of manual labeling, being an important time consuming task to generate the dataset.

4 Classroom Activity Detection

One of the most important things to analyze teaching practices is to observe how does the teacher manage the spoken time during a lesson, which is in fact summarized in a lesson observation metric called *Teacher Talking Time* (TTT). Excessive use of speech by the teacher can lead to students not participating actively as expected during the lesson. In this section we evaluate different approaches to automatically distinguish the teacher's voice from students participation, through the analysis of the audio lesson recording. The basic output of this module is a list of time intervals, with their detected label, which is one of the three values {*teacher*, *multiple*, *student*} already shown in Figure 1.

The problem posed is an audio classification problem, but it can also be tackled using an adapted diarization system [11,15]. So, two different approaches were implemented for comparison:

- **Unsupervised speaker diarization:** answers the question of *who spoke when?*, which means to discriminate all the different speakers in a conversation, and detect the exact segments in which each of them spoke. This is carried out without any prior information about the number of speakers or how their voices sound like.
- **Supervised audio classification:** given enough annotated data with a predefined set of labels {*teacher*, *multiple*, *student*}, a supervised model is trained to predict the label for each audio segment.

Comparing them is relevant because using a pre-trained diarization system does not require any custom training data, thus we consider it a simpler and less costly approach to solve this task. Diarization models are intended to be used on audios with new speakers, whose voices were not known during their training. They do not require fine-tuning over manually annotated data, like the supervised models do.

The unsupervised diarization system is the baseline to which we compare the supervised model. For the latter, we do need training data, which requires a large human effort to annotate the audio recordings. With enough training data, it is expected that the performance for the supervised approach turns out to be better. Thus, one of the key questions to answer is what is the minimum amount of training data needed to surpass the performance of the unsupervised diarization approach.

4.1 Training and Testing Data

As described in Section 3 the dataset considered for this work consists of 25 lessons of 45 minutes each, adding up a total of 19 hours of audio recordings that were manually annotated following the defined protocol. Sections of the recordings with technical issues or that take place before the start of after the end of a lesson were discarded. All of them correspond to situations that would be difficult or impossible to manually annotate and can be considered as outliers and for that reason they were not included in the database. Only the annotated audio segments were used for training and testing, although the rest of the audio was not removed to preserve the context information.

The resulting useful annotated audio totaled 15 hours of recorded lessons, and was split into 50% for training and 50% for testing. This amount of training data is more than enough for the supervised model as we will see next, so the test size was increased for better significance of the comparison results. The data is further divided into groups to analyze how the supervised model performance generalizes for novel teacher voices. For that purpose, the 25 recorded lessons were split into 5 groups, each with voices from 5 different teachers.

To evaluate the amount of annotated data needed, the training set was also divided into 5 splits, without taking into account teacher gender, which were added incrementally to assess model performance. The distribution of labels was approximately: 60% for *teacher*, 30% for *multiple*and 10% for *student*. The splits and groups were designed to keep that distribution as much as possible. This enables to train the models with incremental data in two ways: adding new lessons (i.e. novel teacher's voices), or adding more annotated time for the same lessons.

4.2 Unsupervised Diarization Approach

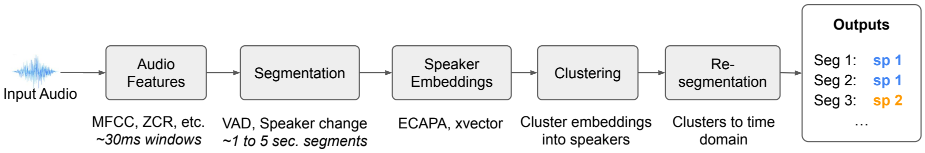


Fig. 2. Diagram of components of a typical diarization system.

Figure 2 shows a typical diarization pipeline, from the input audio waveform, to the output predictions, which consists of a list of segments with their assigned speaker. In our teaching analysis application, diarization does not solve the problem directly, since it does not indicate if any given speaker is the teacher or some of the students. However, we can assume that the teacher is always the most frequent speaker (which was the case in all the lessons recordings that were analyzed for this work). Thus, the speaker detected by the diarization scheme with the largest time across the lesson can be assigned to the teacher label. Please note that this heuristic may fail if the classroom context analyzed is different or if the diarization performance is below a certain quality threshold.

For the implementation, we used the pre-trained *speaker-diarization* pipeline from the *pyannote.audio* toolkit [3]. The module was adapted such that the output speaker with the largest amount of time during the lesson was assigned to label *teacher*. All other speakers were assigned to label *student*. Finally, any audio segment that is not silence neither assigned to *teacher* or *student*, is set to *multiple*.

4.3 Supervised Audio Classification

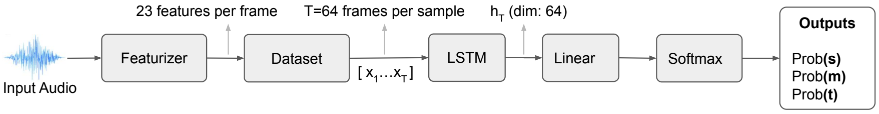


Fig. 3. Diagram of the LSTM-based pipeline for Classroom Activity Detection.

Figure 3 shows the basic blocks of the supervised audio classification pipeline. The key block is an LSTM network with two layers, where the hidden state of the first layer is used as the input to the second layer. The audio features were extracted using overlapping sliding windows that are 30ms long and a hop of 15ms. They consist of 12 MFCC coefficients as well as other audio specific features such as spectral flatness, centroid, bandwidth, contrast (7-dimensional vector), and signal power. The resulting feature vector has 23 components for each 30ms audio frame and is normalized using the mean and variance from all the available training data.

The input to the LSTM network is a sequence of $T = 64$ consecutive normalized feature vectors. This sequence is considered as one audio sample (length $\simeq$ 1s) to be classified as *teacher*, *student* or *multiple*. The classification is done by selecting the largest score after the final layer, which maps the last hidden state h_T from the LSTM to a scores vector, using a fully-connected linear layer and a softmax block. The fact that the hidden state has dimension 64 like the sequence length T is only a coincidence.

4.4 Experiments and Results

Student classroom participations in primary schools are typically short and with lots of overlapping. Most of them are between 1 and 5 seconds, and the voices are not always clear. From the point of view of the quality managers, the goal is not to know the exact boundaries of each student participation. What they look

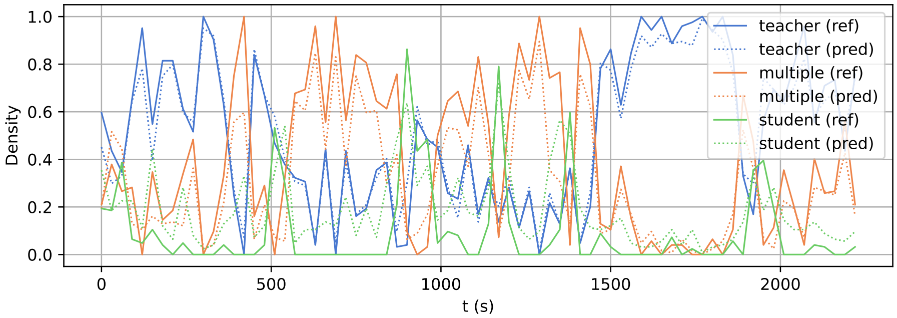


Fig. 4. Example output of the Classroom Activity Detection module, showing the results of the LSTM (supervised) model over 45 minutes of audio.

for is a broader temporal picture of the different moments of the lesson and the participation level. Thus, discussing the optimal resolution with the education technicians, we decided that the most effective value for their purposes was to estimate each label's density on a moving window of 30 seconds. That is to say, to work with label moving averages of that time window, each of them indicating its fraction of speaking time during those 30 seconds.

Figure 4 shows an example of the density estimation for all labels during the 45 minutes of a whole lesson. The LSTM prediction is compared to the reference annotations. As we can see, this visualization allows a user to quickly find peaks of activity in any label, such as moments in which the teacher speaking predominates, or others of high student participation. Given the particular application, an additional metric to the well-known MAE (Mean Absolute Error) was considered. The problem with MAE, as an standard regression metric, is that it evaluates how close the curves are, and not necessarily if they follow the same general shape of peaks and valleys as the reference density. Thus, the other metric considered was the Pearson's correlation coefficient between the predicted and reference density values. Since the correlation is invariant to a vertical shift or scaling in the density estimation, MAE is a good complementary measure to check for systematic over/under estimation problems.

Figure 5 shows the comparison in terms of the correlation coefficient for all labels, and also in terms of MAE for label *teacher* only. In general, the diarization results show a lower accuracy and are less consistent for all the test groups. Test group 1 has very similar correlations between models in label *teacher* (approx. 90%), so to see in more detail what performance this value corresponds to, the estimated densities are shown for both models in Figure 6. Despite the fact that the metrics values are similar and that the densities are actually comparable in terms of accuracy, the main errors in the predictions are in different times of the audio, which is understandable since the underlying considered techniques are very different.

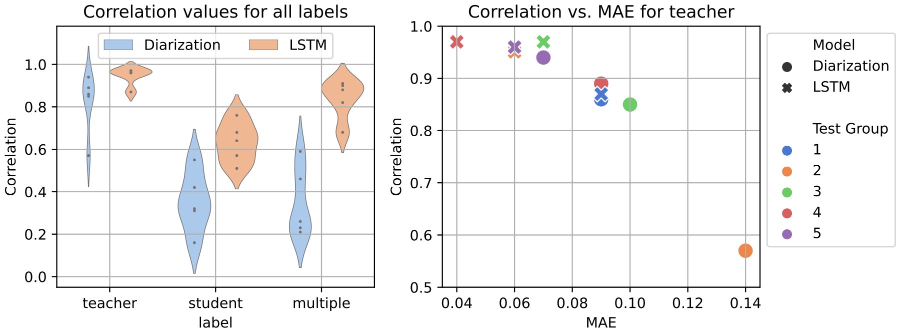


Fig. 5. Comparison between supervised/unsupervised models. **Left:** correlation values for all labels in all test groups (dots inside each violin plot). **Right:** Correlation vs. MAE only for label *teacher*, allowing to see groups individually. Note that test groups 2 and 5 are overlapping for the LSTM model.

For the labels *student* and *multiple*, the diarization system is poorly able to distinguish them with the simple heuristic discussed previously. The confusion is mostly between them, but does not affect the performance for the detection of label *teacher*, where the current approach of taking the most frequent speaker seems successful.

In addition to identifying the moments of teacher and students participation, quality managers also want to have the total time that the teacher or the students spoke, the aforementioned Teacher Talking Time (TTT). For this work, the total estimation error was measured on each of the five test groups, and is shown in Figure 7. It can be seen that it is possible to estimate the TTT with errors below 5 minutes using both approaches. The supervised LSTM model is more precise in the total time for all labels, but it is important to note that for the *student* participation, the average total duration is only 4.5 minutes per lesson (only 10% of the time as mentioned before), so an overestimation of 5 minutes is very significant in this case.

It is also worth noting that the errors in labels *multiple* and *student* are complimentary, so considering them together as one single category reduces the error substantially. In fact, the error distribution becomes complimentary to the *teacher* label, for which the estimation is very precise. In summary, the total time estimation works well to estimate TTT and the overall student participation, but it should not be used to distinguish individual student participation from group work.

After the baseline performance comparison between both approaches, we want now to address the question of what is the amount of data needed to train the supervised classifier, such that it surpasses the performance of the unsupervised diarization pipeline, which does not require any training data. Figure 8 shows the performance improvement as new data groups are added to the training dataset of the LSTM classifier. The correlation is measured only on the test

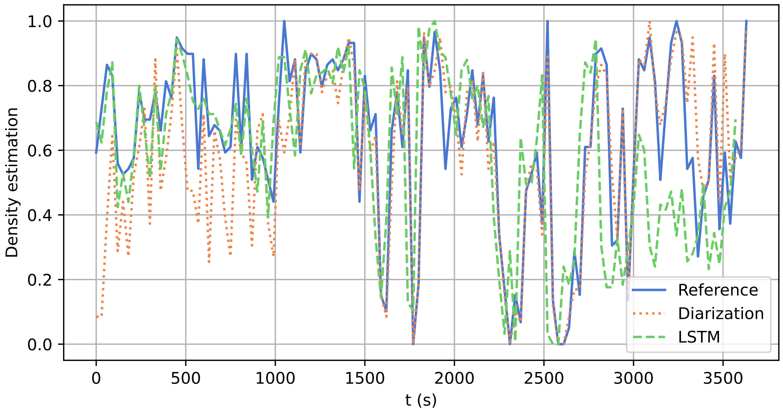


Fig. 6. Comparison of label *teacher* densities over **test group 1**. This is the group where the diarization and LSTM have the most similar metrics.

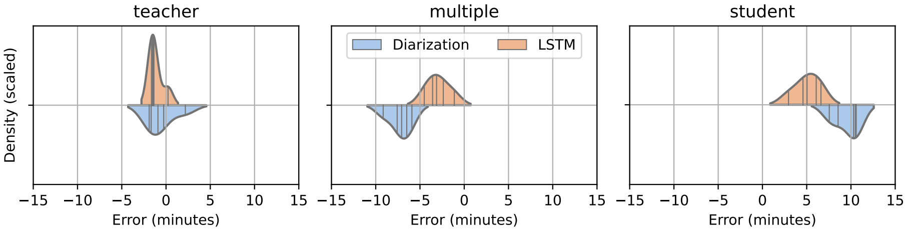


Fig. 7. Distribution of the total time estimation error for each label, for both models. The error is scaled to the duration of a standard lesson of 45 minutes. Error for the teacher time corresponds to the TTT.

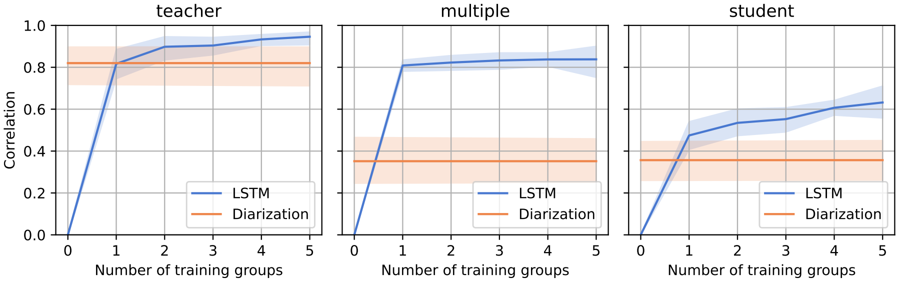


Fig. 8. Comparison between the unsupervised diarization model vs. the supervised LSTM approach, while adding new groups of audios (i.e., adding new lessons and teachers) to the supervised model. The colored spans around each curve represent the confidence interval of 95% of each metric, as measured on all 5 test groups.

groups that were not present in the training sets, except for the last point which includes all the five groups (but always keeping separate data to train and test). This means that except for the last data point, the classifier is being evaluated on new -unknown- teacher and student voices. These results could be used to estimate how the classifier will generalize and perform on new voices.

With respect to adding more annotated time from the same lessons, the detection of *teacher* and *multiple* labels reaches the plateau with the first training split (less than two hours of annotated data). The only label which keeps slightly improving in this case is the *student*, which makes sense considering the low proportion of samples that this label has. Hence we deduce that trying to annotate the whole lesson is not necessary, but instead the efforts should be focused on times with high student participation.

5 Additional Tools for English Lessons Analysis

For the particular case of English lessons, in addition to the speaker detection detailed in the previous section, some additional tools were developed to help in the analysis carried out by the quality managers. In the next sub-sections we present the approaches followed for language identification and key phrases detection. We end up this section introducing the user interface developed for the education quality managers, to support their lessons analysis work.

5.1 Language detection

During English lessons, the best teacher practice is to encourage the use of English language as much as possible. Spanish usage (mother language in Uruguay) is only justified when they have technical issues due to the videoconference system. Thus, a language detection module is very helpful for the quality managers to analyze the English usage during the lesson. This tool enables to automatically alert when excessive Spanish usage is detected in a lesson.

As introduced in Section 2, the Whisper model [12] includes a language detection module prior to speech transcription. In this work, we exploit such functionality to detect the spoken language during each segment of the lesson audio recording. Before the language detection, the audio recording is segmented using a voice activity detection (VAD) module. For this purpose we used the same implementation from *pyannote.audio* included in the diarization pipeline shown in Section 4.

The prediction is based on a probability vector for the different possible languages. These probabilities are computed for each speech segment based on all the languages included in the model. Each value represents the algorithm confidence to assign each language to the audio segment. Thus, we decide the language spoken on each segment, as the one with the largest probability. Figure 9 illustrates the pipeline developed, where the raw audio recording is segmented with the VAD and each segment is classified as English or Spanish, according to the output probability vector for each speech audio segment. Based on the probabilities, it is also possible to define a threshold when both values are close to each other, in order to indicate borderline cases to be further analyzed by manual inspection.

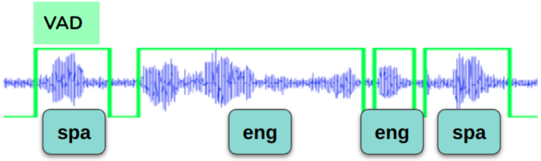


Fig. 9. Language detection example, where the blue signal is the raw audio waveform and the green line is the VAD output. Each audio segment is classified as English or Spanish, based on the maximum confidence values of the Whisper language model.

For the evaluation of this language detection pipeline, another manual labeling was also necessary. A small dataset of four lessons was selected, totaling three hours of recorded audio labeled. Based on this data, the corresponding confusion matrices were generated (shown in Figure 10), considering the amount of time for each language. The results show much better accuracy for English language (eng) detection, with almost perfect accuracy, unlike the one observed for the less used Spanish language (spa). In a deeper look at every single lesson recording, it was noticed that the latter improved when Spanish usage is greater. For example, the accuracy for the audio recording where more Spanish is spoken (19% of the total lesson) reaches 90%.

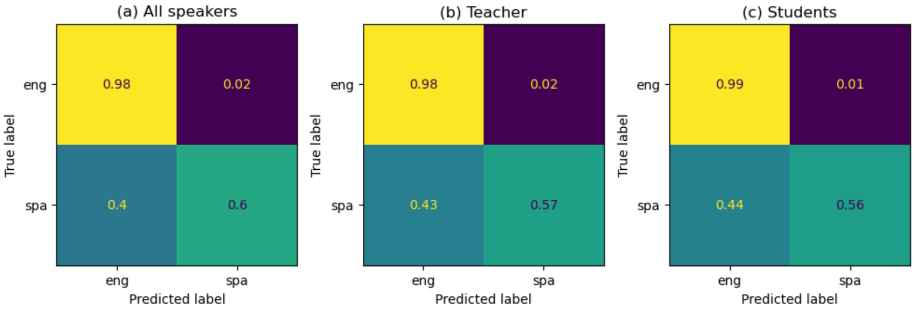
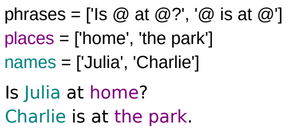


Fig. 10. Language detection confusion matrices for different speakers.

This observations verify that the proposed approach can still be effective for the desired application, which is to detect significant lesson segments where the language usage does not correspond to the best teaching practice of English usage. The performance drops detected, correspond to cases where very little



```
phrases = ['Is @ at @?', '@ is at @']
places = ['home', 'the park']
names = ['Julia', 'Charlie']

Is Julia at home?
Charlie is at the park.
```

Fig. 11. Phrase detection example with *wildcards*. Each "@" can only be replaced with a word belonging to one of the predefined lists.

Spanish usage occurs, so it is not a relevant situation for the teaching analysis application. Moreover, combining the result of this module with the one presented in the previous section, it is possible to identify if the language misusage corresponds to the teacher or the students.

5.2 Key phrases matching

Key phrases detection enables to find out if the expected grammar and vocabulary of the corresponding Unit is used during the lesson and to assess the number of times that are repeated. This is particularly important for English lessons, since each Unit has predefined learning exercises. Thus, one relevant thing that the quality managers seek to verify is if the appropriate vocabulary is trained during each lesson, analyzing the speech of both the teacher and the students.

The implementation of this module was also based on the Whisper model [12], but in this case using the speech-to-text output. In addition to the lesson transcription, another required input is the list of strings indicating the grammar and vocabulary that should be detected for a particular Unit. Finally, the goal is to search all the key phrases occurrences during the lesson, indicating for each of them the corresponding times in which they were detected. Since the transcription is not always perfect, a dynamic programming approach was implemented, using the Levenshtein distance to find the most similar matches.

As phrases may have variants, we integrated *wildcards* to deal with them, taking into account the minimum distance over all possible words for each wildcard. Figure 11 shows an example where two lists of places and characters have a predefined vocabulary that can be used. The purpose of wildcards is to allow flexibility in the vocabulary detection used in the lesson. The module detects the times at which the lesson transcription matches an entry from a set of predefined phrases defined by their grammar and vocabulary. The phrases are organized hierarchically so that each phrase corresponds to an educational Unit, allowing the module to produce a report indicating which and when the different Units have been covered in a lesson.

5.3 User Interface for Education Technicians

The final application requires the integration of all the previously described modules in a friendly user interface to be used by the quality managers. This is achieved through a simple web interface, where the education technicians select the lesson recording to be analyzed, enters some basic information about the teacher and the students group, in addition to the options of the required report (e.g. if language usage analysis is necessary or the word list for the key phrases detection).

Combining the results of the different modules (i.e. speaker identification, language usage and key phrases matching) a PDF report is automatically generated with all the relevant metrics summary for the particular lesson and the teaching practice observed. An output video is also generated, which enables to easily navigate through the recording, going directly to relevant excerpts for the quality managers, such as moments of high interaction with students, excessive use of Spanish or vocabulary usage associated with specific content from a course Unit. With the developed tool the teaching evaluation process is enhanced, supporting the quality managers with objective data which enables a faster and more detailed lesson analysis.

6 Conclusions and Further Work

In this work we present different machine learning modules integrated in a tool for education technicians who work on teaching practices analysis and evaluation. Based on lesson recordings for a particular remote teaching scenario, we generated a manually annotated dataset which serves both algorithm development and evaluation.

The first module deals with the speaker identification problem. The goal is to analyze how does the teacher manage the speaking time during the lesson and how much participation do the students have. For this purpose we compared two different approaches, an unsupervised diarization system versus a custom-trained LSTM network. The latter showed a better performance as expected, but we also validated that the diarization approach could be enough if we are only interested in the teacher speech.

Two additional modules were presented for English lessons. The first one uses language detection to identify excessive use of Spanish, while the other is focused on key phrases matching associated to particular Unit topics. All the modules were integrated into a web application, in order to help the quality managers with the teaching practices analysis and evaluation.

Further discussions with the quality managers, as soon as the tool is used more intensively, will bring novel requirements to be addressed. For example, we plan to add the detection of predefined recordings played during a lesson which indicate specific activities were covered, such as English listening exercises. We also expect to obtain new statistical information using the presented tools to contribute to research on education studies carried out within Ceibal.

Acknowledgements This research was funded by Agencia Nacional de Investigación e Innovación (ANII) Uruguay, Grant Number FMV_1_2021_1_166660.

References

1. Banegas, D.L.: ELT through videoconferencing in primary schools in Uruguay: first steps. Innovation in Language Learning and Teaching **7**(2), 179–188 (2013)
2. Blunt, P., Haskins, B.: A model for incorporating an automatic speech recognition system in a noisy educational environment. In: 2019 International Multidisciplinary Information Technology and Engineering Conference (IMITEC). pp. 1–7 (2019)
3. Bredin, H., Yin, R., Coria, J.M., Gelly, G., Korshunov, P., Lavechin, M., Fustes, D., Titeux, H., Bouaziz, W., Gill, M.P.: `pyannote.audio`: neural building blocks for speaker diarization. In: IEEE ICASSP (2020)
4. Cosbey, R., Wusterbarth, A., Hutchinson, B.: Deep learning for classroom activity detection from audio. In: IEEE ICASSP. pp. 3727–3731 (2019)
5. Foil, J.: Language identification using noisy speech. In: IEEE ICASSP. vol. 11, pp. 861–864 (1986)
6. Guimarães, L.M., da Silva Lima, R.: A systematic literature review of classroom observation protocols and their adequacy for engineering education in active learning environments. European Journal of Engineering Education **46**(6), 908–930 (2021)
7. Kaplan, G.: Innovations in education: Remote teaching. British Council, London, UK (2019)
8. Martinez, J., Perez, H., Escamilla, E., Suzuki, M.M.: Speaker recognition using mel frequency cepstral coefficients (MFCC) and vector quantization (VQ) techniques. In: 22nd International Conference on Electrical Communications and Computers. pp. 248–251 (2012)
9. Millman, J., Darling-Hammond, L.: The New Handbook of Teacher Evaluation: Assessing Elementary and Secondary School Teachers. Corwin Press Inc., SAGE Publications (1990)
10. Owens, M., Seidel, S., Wong, M., Tanner, K.: Classroom sound can be used to classify teaching practices in college science courses. PNAS Psychological and Cognitive Sciences **114**(12), 3035–3090 (03 2017)
11. 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 **72**, 101317 (2022)
12. Radford, A., Kim, J.W., Xu, T., Brockman, G., McLeavey, C., Sutskever, I.: Robust speech recognition via large-scale weak supervision. arXiv CoRR **abs/2212.04356** (2022)
13. 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: 11th LAK Conference. p. 132–140 (2021)
14. Slyman, E., Daw, C., Skrabut, M., Usenko, A., Hutchinson, B.: Fine-grained classroom activity detection from audio with neural networks. arXiv CoRR **abs/2107.14369** (2021)
15. Wang, Q., Downey, C., Wan, L., Mansfield, P.A., Moreno, I.L.: Speaker diarization with LSTM. In: IEEE ICASSP. pp. 5239–5243 (2018)
16. Zissman, M.A., Berkling, K.M.: Automatic language identification. Speech Communication **35**(1), 115–124 (2001)
17. Zylich, B., Whitehill, J.: Noise-robust key-phrase detectors for automated classroom feedback. In: IEEE ICASSP. pp. 9215–9219 (2020)