Classroom Activity Detection in Noisy Preschool Environments with Audio Analysis

Siddhartha CV
Department of Electrical Engineering
Indian Institute of Technology Bombay
Mumbai, India
siddharthsh2@gmail.com

Preeti Rao
Department of Electrical Engineering
Indian Institute of Technology Bombay
Mumbai, India
prao@ee.iitb.ac.in

Rajbabu Velmurugan
Department of Electrical Engineering
Indian Institute of Technology Bombay
Mumbai, India
rajbabu@ee.iitb.ac.in

***Abstract*—Pre-school teaching demands several diverse methodologies making the quantification of classroom activity very useful to reflect on and refine teaching practices. Therefore, automating the monitoring process with unobtrusive audio recording and analysis is viewed as a simple and scalable solution. We present a system for performing automatic classroom activity detection from audio using a new audio corpus obtained from preschool classrooms in rural India. We define the sound classes of interest in estimating the level of teacher engagement. To address the challenge of training data scarcity, we investigate the combination of pre-trained models with unsupervised homogeneity-based segmentation to realise a multilabel classification scheme. We discuss the performance of the classifier and its utility in the context of the classroom activity monitoring task.**

***Index Terms*—classroom activity detection, audio diarization**

I. Introduction

Early childhood education studies have shown that incorporating student-centric active learning strategies in the classroom can improve the student retention rate, encourage class participation, and boost the overall effectiveness of instruction. Instructors and curriculum designers hence need to quantify classroom activity in order to review, reflect and revise these techniques, based on their impact on student learning and engagement. This can be achieved through manual monitoring of classroom sessions, which unfortunately is human resource-intensive. Classroom Activity Detection (CAD) aims to reliably identify speakers and events (teachers, students, group activities, etc.) through the analyses of audio recordings. The challenges include the extensive presence of overlapping talk between teachers and students, with common occurrence of short speaker turns making it hard even for a human listener to assign roles. Further, classroom recordings are typically noisy and may be of non-uniform quality due to the varying distances of the different speakers from the recording device. Finally, a CAD model also needs to scale well to unseen speakers.

CAD, although a relatively new problem, has attracted much research interest. In early work, a modified proprietary wearable device, "LENA," [1] was used in K-12 classrooms, the audio features from which were mapped to classroom activities: "teacher lecturing," "whole class discussion" and "group work" at a 30 second temporal granularity using a random forest classifier [2]. A set of classroom activities were automatically detected using instructor-worn close talking noise-cancelled microphone data collected from middle school classrooms, with a Naive-Bayes classifier on assorted prosodic, NLP and acoustic features [3].

The Decibel Analysis for Research in Teaching (DART) tool [4] enabled large scale, automatic annotation of classroom activity in college-level lecture audio using portable audio recording devices. DART uses binary decision trees on statistics of the energy over local windows of the audio signal to perform multi-class classification on a 4-way label set: "single voice," "multi voice," "no voice" and "other". Substantial reductions in frame-level error over DART were demonstrated with a gated recurrent unit (GRU) neural network using Mel-filterbank features, for new class sessions from both previously seen and unseen instructors [5].

A 9-way classification (also simplified to 4-way and 5-way tasks) is performed using various neural network architectures on a university classroom corpus with a duration of 58.7 hrs [6]. There, a bi-directional GRU model is trained on Mel-filterbank, OpenSmile and PASE+ self supervised acoustic features, obtaining frame-level error rates of 6.2 %, 7.7 % and 28.0 % when generalising to unseen instructors for the 4-way, 5-way, and 9-way classification tasks, respectively. Although the 5-way task is similar to that considered in this paper, we deal with data that exhibits significant polyphony from the overlapping labels. Thus this is the first reported work on the following realistic and challenging scenarios.

1) **Pre-school classrooms and multilabel classification:** We work on audio recordings from pre-schools, which are noisy and have high activity and speaker-overlaps, requiring a multilabel classification scheme, opposed to earlier works that generally focused on university-level STEM classrooms.
2) **Choice of labels:** Due to the nature of our classrooms, we introduce labels such as 'Multiple Speakers' and 'Striking' to represent disruptions during class time, which are important to detect in view of interest in class time use and student engagement.
3) **Dataset size:** We have access to a much smaller dataset compared to earlier work in CAD, with a total duration of 3 hrs 39 min, limiting us to lighter models in contrast

to RNNs and GRUs that show state-of-the art results for similar tasks.

We address these challenges by combining two relatively independent stages of audio processing. One stage estimates the classes at frame-level (i.e. at a fine time resolution of 100 ms) and the other detects activity boundaries via unsupervised segmentation based on homogeneity. Frame multilabels are aggregated within the predicted activity boundaries to obtain the required segmentation at the higher time scales that will eventually help summarise the activities in the classroom. We propose a novel architecture that uses three different features (log-Mel spectrogram, pre-trained embeddings from Wav2Vec2 and OpenL3), motivated by our limited dataset that also incorporates diverse speech and non-speech classes, to perform a frame-wise multilabel classification. We also handle class-imbalance by proposing a class-wise dynamic thresholding mechanism to obtain fine-grained classification of the audio frames.

II. Dataset and Task

A. Dataset

The dataset used in this work was provided by EarlySpark [7], an organisation devoted to creating optimal learning environments for young children. They partner with schools and train teachers to enhance formative childhood education. The dataset is of a total duration of 3 hrs 39 min, containing recordings from 24 classroom sessions with 8 different instructors (7 female, 1 male). The recordings, from preschool classroom sessions using an audio recorder placed in the room, vary in duration between 4 to 22 min with a mean session duration of 9 min 15 sec, recorded at 16 kHz sampling rate.

The recordings were manually annotated by a single rater using 6 labels as in Table I. The annotation is carried out using `Audacity`, such that up to three labels can co-occur. The relative overall durations of the labels in the dataset are shown in Figure 1, with the total duration depicted in the pie chart being 4 hrs 17 min. This is higher than the actual dataset duration due to multiple instances of label overlap. The high probability of having more than one label for a given frame (16% frames with event overlaps) motivates the multilabel classification scheme, which has not been attempted previously for the CAD task. Apart from high overlap, we also observe

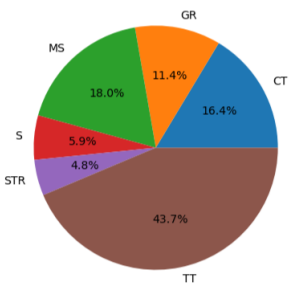


Fig. 1: Distribution of labels across the dataset, with a total duration of 4 hrs 17 min including overlaps

TABLE I: Labels used during annotation along with their description

Event	Label	Description
Child Talking	CT	Child being the dominant speaker
Teacher Talking	TT	Teacher being the dominant speaker
Group Response	GR	Children responding in chorus
Multiple Speakers	MS	Many simultaneous unrelated voices
Striking	STR	Occurrence of hitting, clapping, jumping, stomping, banging the table etc.
Silence	S	Region with no speech activity and striking events, but could have ambient background noise

frequent turn-taking, with majority of activities/speaker turns having 0.5 to 2 s duration.

We also distinguish between speech and non-speech classes, with 'CT', 'TT', 'GR' and 'MS' forming the speech classes, while 'S' and 'STR' are considered non-speech classes. The labels 'MS' and 'STR' were introduced to capture disruptions of normal class activity due to the nature of children at that age, increasing the complexity of the problem due to acoustic overlaps with labels like 'CT' and 'GR'.

Of the 24 recorded class sessions from 8 instructors, we create a `Training` dataset of 19 sessions and two `Test` datasets. `Test 1` comprises of 3 sessions (≈24 min) with the instructors overlapping with the `Training` (i.e. seen speakers) `Test 2` comprises of 2 sessions (≈15 min) from instructors not seen during training, including the single male instructor, to see how the model generalizes to unseen speakers. The `Validation` set was created by randomly selecting 10% of samples from the `Training` set and not used in training.

B. Task and Evaluation Criteria

For our classification task, we window the audio signal to generate frames with 500 ms context and 100 ms hop. Let F be the sequence of frames in a classroom recording, i.e., $F = \{f_i\}_{i=1}^{N}$ where f_i denotes the i^{th} frame and N is the total number of frames. Let Y be the target label sequence, i.e., $Y = \{y_i\}_{i=1}^{N}$ where each y_i is a 6-dimensional multilabel vector with binary entries $\{y_{ij} = 0 \text{ or } 1\}_{j=1}^{6}$ corresponding to the absence/presence of a classroom activity type in the frame based on manual annotation, in the order ['CT','GR','MS','S','STR','TT']. We define the classification problem as follows:

Given a frame f_i, and input features x_i, our goal is to find the most probable multilabel, $\hat{y}_i$ for the frame as follows:

$$\hat{y}_i = \underset{y \; in \; \mathbb{Y}}{\arg\max} \; P(y|x_i) \tag{1}$$

where $\mathbb{Y}$ is the collection of all possible multilabels and $\hat{y}_i$ is the predicted multilabel for frame f_i. A standard approach while dealing with multilabel targets is by minimising binary cross entropy for each class and using sigmoid activation at the output node to obtain class-wise probabilities. We evaluate our models at a frame-level using the standard metrics: macro f1-score (MF) and weighted f1-score (WF), and also report class-wise f1-scores (CF).

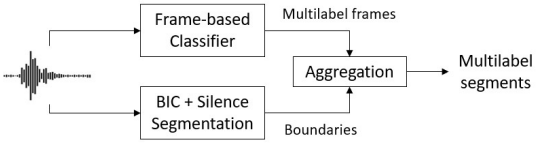


Fig. 2: Block diagram for the proposed system

III. Methods

Our proposed system consists of two parallel stages and an aggregation step as shown in Figure 2. One stage is a classifier that uses features extracted from raw audio to generate frame-wise multilabel predictions. The other stage is a Bayesian Information Criterion (BIC) based based segmentation of the audio to obtain the class or event boundaries.

A. Frame-level classification

The following steps are involved.

1) Feature extraction: Apart from raw spectral features linked to timbre, we consider feature concatenation with two available pre-trained audio embeddings that can potentially alleviate the scarce data problem.

- **log Mel-filterbank:** Log Mel-filterbank features have been widely used in applications like speaker verification [8] and audio event detection [9] [10]. We extract 128 log-scaled Mel-filterbank features using the Python audio processing and analysis package `librosa` [11] with standard 25 ms windows at 20 ms frame rate.
- **Wav2Vec2:** The `Wav2Vec2-Base-960h` model generates speaker representations from raw audio, fine-tuned on the Librispeech [12] dataset containing 16 kHz sampled speech from 2338 speakers. Self-supervised learning is used to learn a set of speech units and describe the speech audio sequence. Given its use in speaker recognition tasks [13] [14], these features are included to help distinguish speech events. The model consists of 12 transformer blocks that help analyze a larger context. 768 dimensional features are extracted from the last block using a 25 ms window, and a 20 ms hop.
- **OpenL3:** OpenL3 is an open-source Python library for computing deep audio and image embeddings using a self-supervised learning method of audio-visual correspondence [15] on AudioSet videos [16]. By employing the 'env' content model trained on environmental videos, we look to better identify non-speech classes, given the datasets' acoustic diversity. We extract 512 dimensional embeddings using the default frame definitions - 1s context and a 100 ms hop.

The use of features with a larger context has been effective in CAD tasks [5] [14]. To achieve this, as well as synchony across the 3 feature streams, we concatenate the log Mel-filterbank features over 500 ms at 100 ms hop. Similarly, the Wav2Vec2 vectors are average pooled before the feature concatenation.

2) Model Architecture: As shown in Figure 3, the mel filterbank features are processed with a CNN of three 2D convolutional blocks (ConvBlock). Each convolutional block

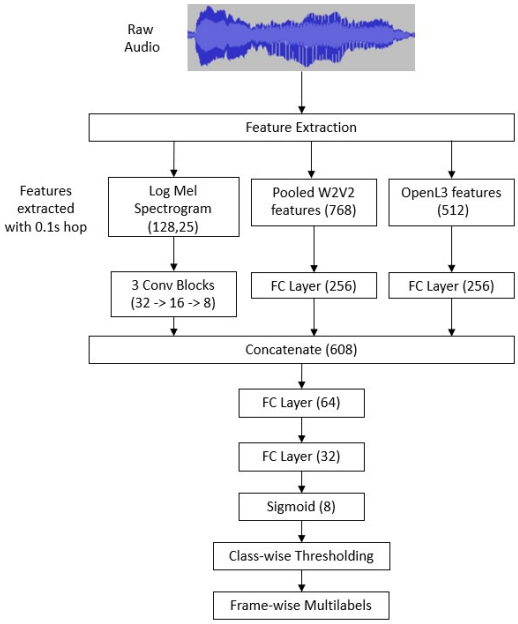


Fig. 3: 3-branch architecture of the frame-based classifier

consists of one 2D convolutional layer with a kernel size of (3x3), stride of (1, 1), without any padding. After each convolutional layer, we use batch normalization, ReLU activation and max pooling with a pool size of (4×2) and stride (4, 2). We use three ConvBlocks having a decreasing number of convolutional filters: 32, 16 and 8. The output from the last ConvBlock is flattened before concatenation, resulting in a 96 dimensional vector. Wav2Vec2 and OpenL3 inputs are fed into separate dense layers with 256 neurons, and the output of all three branches are concatenated to give rise to a 608 dimensional vector. The resulting combined parameter vector is used as input for a block of dense layers, with two fully connected layers having 64 and 32 neurons. Finally, a sigmoid layer is used to extract probabilities of each activity type for all frames.

For training, Adam optimizer [17] is used with a learning rate of $5e^{-5}$ to minimise Binary Cross Entropy (BCE) loss at the output unit of each class on the multilabel targets. Batch size is set to 64 and the model is trained for up to 20 epochs, with early stopping used on the `Validation` set and patience set to 7 epochs.

3) Class-wise Thresholding: From the classifier, we extract frame-wise prediction vectors $\{p_i\}_{i=1}^{N}$, containing elements $\{p_{ij}\}_{j=1}^{6}$, each representing the probability (between 0 and 1) of the presence of the activity types, ['CT','GR','MS', 'S','STR','TT'] in that frame. In order to take care of the class imbalance present in our dataset, we threshold classifier predictions, along the lines of frame-wise contour-based dynamic thresholding introduced by [18] for polyphonic acoustic event detection, with slight modifications. First, a probability contour is derived by plotting for each frame the highest output probability for that frame across the 6 event labels, called the confidence for that frame, $C_i = max(p_{i1}, p_{i2}, p_{i3}, p_{i4}, p_{i5}, p_{i6})$. However, instead of using a single global threshold for all

labels, we use a vector α, with different global thresholds for each of the 6 labels. α_j for each class j is obtained by taking the median of the sigmoid output probabilities from the classifier on `Training` data samples where class j was active. C_i is included to take into account the level of polyphony present in a given frame, while the α vector is used to mitigate the class imbalance present in our dataset.

The threshold for frame index i can be defined as $T_i = \{t_{ij}\}_{j=1}^{6}$, where $t_{ij} = \alpha_j * C_i$. The decision of which labels are active in a given frame, generating multilabel predictions $\hat{Y} = \{\hat{y}_i\}_{i=1}^{N}$, is made using the above identified thresholds as follows:

$$\{\hat{y}_{ij} = 1 \; if \; p_{ij} > t_{ij}\}_{j=1}^{6} \tag{2}$$

From our definition of the different classroom events and on analyzing manual annotations to determine the extent of polyphony, we restrict our predictions to only three classes, the top two speech classes and the non-speech class with the highest probability, which are then compared with the frame-wise thresholds to generate multilabel predictions.

B. Segment Boundary Detection

The segment boundary detection step involves using a segmentation algorithm that combines Bayesian Information Criterion (BIC) based segmentation using MFCCs [19] and detection of pauses and silence based on variation of RMS energy [20]. Since pauses are often present at speaker changes, they are marked by lower RMSE compared to surrounding regions, and can hence be used to mark boundary points. However, speaker changes can also take place where there are no significant changes in RMSE, which are better identified by the BIC based algorithm. These two standalone segmentation approaches are combined to ensure ground truth boundaries are not missed. Thus, while evaluating their performance we look to maximize recall, while maintaining significant precision. The unsupervised segmentation algorithm achieves recall > 0.9 and precision > 0.5 across files with a 0.5 s boundary tolerance, tuned using four sessions from the `Training` dataset.

C. Post Processing via Aggregation

The frame-wise multilabels obtained as output from the classifier (after thresholding), are at too fine a granularity for easy analysis during downstream applications. Thus, there is a need to convert these predictions into a form that best resembles manual annotation. Also, frames close to segment boundaries capture information from neighbouring events, and classifier predictions for those frames (as described in Section III-A) tend to be weaker in confidence. These predictions can be refined by aggregating them with non-boundary frames in their vicinity, based on automatically detected speaker change boundaries as described in Section III-B. This aggregation is performed so that all frames present between two speaker change points have the same multilabel, removing any irregularities in model predictions, and generating segment-wise multilabels that have a lower granularity, better suited for analysis. It is performed as follows.

Given a set of predicted boundaries $B = \{b_k\}_{k=1}^{M}$, where $b_0 = 0$ and M is the total number of boundaries predicted, a segment s_k is defined as the region between boundaries b_{k-1} and b_k. For a segment s_k containing frames $\{f_i\}_{i=p}^{q}$, we select the most common predicted multilabel for frames within s_k as the multilabel for all the frames in s_k, i.e., $\{y_i^c\}_{i=p}^{q} = mode(\{\hat{y}_i\}_{i=p}^{q})$. Hence, this method allows us to move from fine- to coarse-grained predictions, to generate a label track similar to that of manual annotation.

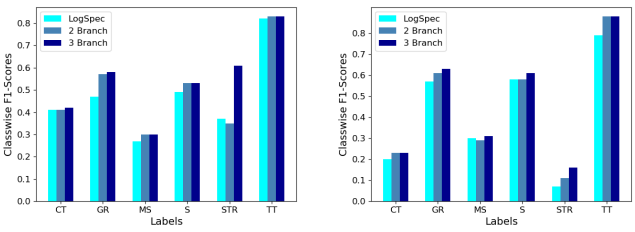
IV. Experiments[1]

In order to assess the change in classifier performance on including pre-trained embeddings, a model consisting of only the log Mel-spectrogram branch is initially optimized (LogSpec model), to which the Wav2Vec2 branch is added and tuned (2 Branch model). Finally, we include the OpenL3 branch, resulting in the 3 Branch architecture shown in Figure 3. The models are trained, and class-wise f1-scores are obtained from the multilabel predictions for both test sets, shown in Figure 4a for `Test 1` and Figure 4b for `Test 2`.

Model Comparison: We see that all three models perform best on the dominant 'TT' class, with the f1-scores for the other classes lying in the range of 0.3-0.6 for the best model. On including the Wav2Vec2 branch, we see a marked performance improvement for the speech classes. We get better class-wise f1-scores for 'S' and 'STR' on introducing OpenL3 features across both test sets. These justify our motivation to include them. We also observe that there is some degradation in performance from `Test 1` (unseen class sessions from previously seen instructors) to `Test 2` (unseen class sessions from unseen instructors), however the model generalizes comparably. Due to relatively fewer cases of 'CT' and 'STR' present in `Test 2` compared to the original distribution (Figure 1), we see a noticeable drop in performance, while for other classes the models' performance remains at par with that in `Test 1`.

Classifier Performance: In order to understand which classes are typically confused by the best model, we look at the confusion matrix generated on the predictions of both test sets, when only a single class is active in the ground truth.

[1]Annotated data and code are available at: https://gitlab.com/siddharthsh2/classroom-activity-detection-in-noisy-preschool-environments.git



(a) CF on `Test 1` (b) CF on `Test 2`

Fig. 4: Class-wise f1-score comparison of the 3 models

TABLE II: Confusion matrix when only a single class is active. Shaded boxes indicate significant mis-classification

Ground Truth \ Predicted	CT	GR	MS	S	STR	TT
CT	925	225	822	574	79	517
GR	267	1873	631	120	170	242
MS	173	551	595	80	15	216
S	22	1	16	128	3	21
STR	23	20	15	32	127	13
TT	1390	220	1024	1067	303	7754

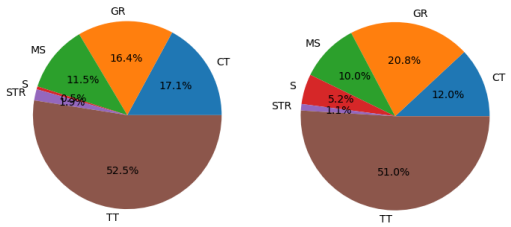
The major cases of mis-classification are observed when we look at classes 'CT' and 'MS', which have lower f1-scores compared to the other speech classes. From Table II, 'CT' is heavily mis-classified as 'MS', possibly due the noisy nature of the classrooms, where children are gossiping even in the presence of a dominant child speaker, which could be labeled as either 'CT' or 'MS', depending on the annotator. 'MS' is majorly mis-classified as 'GR', since both classes have children speaking together, the only difference being whether or not they are in sync/ talking in chorus. Occasionally, 'CT' and 'MS' are also mis-classified as 'TT', due to the models' inherent bias towards the majority class.

We also compare frame-level classification performance with and without dynamic thresholding and post-processing techniques, on the 3 branch model outputs, which can be seen in Table III. First, we note the performance on using a single global threshold for all classes, obtained by tuning on the `Validation` set (threshold set to 0.3 for `Test 1` and 0.25 for `Test 2`). Then, we look at the change in MF and WF brought about using the dynamic thresholding explained in Section III-A3. Finally, we implement post-processing (PP), both with boundaries automatically detected at high recall from the segmentation task (HR boundaries) and the ground truth boundaries from manual annotation (GT boundaries) to know the extent of performance improvement that could be achieved given perfect segmentation.

From Table III, we observe that using a dynamic threshold leads to an increase in macro and weighted f1-scores across both `Test` sets, since our strategy allows different classes to have different thresholds based on both confidence of model prediction (easier to predict classes to have higher thresholds and vice-versa) and estimated frame-level polyphony. Next, we observe that post-processing with HR boundaries leads to an improvement for `Test 1` but not for `Test 2`, which may be attributed to comparatively weaker predictions for unseen teachers. However, on using GT boundaries, we see a significant improvement in scores with an overall 11 % increase

TABLE III: Frame-level F-scores with and without dynamic thresholding and post-processing techniques on 3 branch model

Technique	Test 1		Test 2	
	MF	WF	MF	WF
Single Threshold	0.52	0.63	0.46	0.63
Class-wise Thresholding	0.54	0.64	0.47	0.64
With PP (HR Boundaries)	0.54	0.65	0.45	0.62
With PP (GT Boundaries)	0.60	0.70	0.52	0.68



(a) Ground Truth (b) Predicted

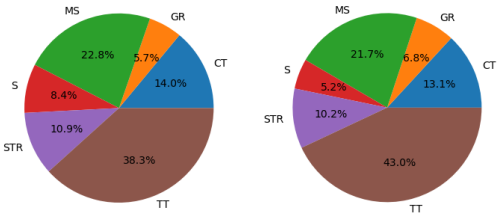
Fig. 5: Classroom activity distribution during a 5 min snippet with high engagement in a classroom session

in macro f1-score and an 8 % increase in weighted f1-score, suggesting some scope for improvement of our segmentation algorithm, and demonstrating the potential of this technique to provide better predictions even when a models' performance saturates.

V. Downstream Application

To demonstrate the usefulness of the system described in Section III, even at its current performance levels, we apply it in a classroom monitoring exercise. First, we analyze a classroom audio recording over a 5 min interval to obtain the distribution of various classes in the generated label track. This provides a snapshot of the extent of participation or classroom activity taking place. Based on our selection of classes, it is possible to identify regions of high engagement, purely teacher instruction, or those of low engagement.

In Figure 5, we see that the 5 min snippet is dominated by 3 classes - 'TT', 'GR' and 'CT', from which we can conclude that this part of the session was one of high engagement. Such distributions are typically found when a turn-taking behaviour is observed, i.e., teacher and child(ren) speak alternatively, suggesting question-answer sessions, teacher asking students to repeat a particular word/phrase after her, or clarification of doubts



(a) Ground Truth (b) Predicted

Fig. 6: Classroom activity distribution during a 5 min snippet with low engagement in a classroom session

Alternatively, Figure 6 shows a distribution from a 5 min snippet where a significant portion is captured by the 'MS' and 'STR' classes, suggesting some disruption from children during that time, repeated occurrences of which in a particular session could indicate ineffective use of class time by the instructor. In both cases, we observe that the predicted label distribution comes close to the ground truth in terms of

capturing the environment of the classroom in that period, making it a powerful tool for administrators.

As a by-product of our work, an analysis tool was developed to extract the occurrences of any chosen label in a recording, and synthesizing a new track while retaining information about the original timestamps. This facilitates the rapid listening based analysis of a classroom recording by an administrator. For instance, if they wish to check whether the instructor adheres to the curriculum, only regions corresponding to 'TT' in the label track can be played out. Given that 'TT' generally occurs only 44 % of the time (Figure 1), the generated label track along with the analysis tool eliminates the need to go over the entire recording. Similarly, 'CT' and 'GR' regions can be isolated to estimate the amount of positive engagement of the class with the instructor.

VI. Conclusions

We proposed a system for classroom activity detection that achieves a competitive performance with noisy classroom data that also has significant speaker/event overlaps. We see significant benefits from combining pre-trained embeddings with conventional acoustic features at the frame level. The chosen pre-trained embeddings compensate for the shortage of labeled speech and general audio data respectively. Introduction of class-wise dynamic thresholding leads to better multilabel predictions by incorporating information about frame-level polyphony and the class imbalance inherent to the dataset.

Integrating high recall segmentation boundaries with the frame-wise classifier outputs allows us to suitably refine predictions using the knowledge of speaker change points. The true potential of this technique is evident when using ground truth (i.e. oracle) boundaries for aggregation indicating avenues for future work on better segmentation algorithms. We note that aggregation using homogeneity based speaker change boundaries is applicable in general to sound event detection (SED) or speaker diarization, for refining frame-level predictions. Other possibilities for future work include exploiting further speech-specific features such as fundamental frequency similar to the OpenSmile features [21] used to better separate speech classes [6] and the use of techniques like mixup [22] for data augmentation.

VII. Acknowledgements

The authors gratefully acknowledge the EarlySpark team, especially Smita Deshpande, Anuja Burns and Lav Varshney, for their critical role in defining this project and providing valuable feedback throughout the course of the work presented here.

References

[1] M. Ford, C. T. Baer, D. Xu, U. Yapanel, and S. Gray, "The LENA language environment analysis system," 2008.

[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] 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.

[4] M. T. Owens, S. B. Seidel, M. Wong, T. E. Bejines, S. Lietz, J. R. Perez, S. Sit, Z.-S. Subedar, G. N. Acker, S. F. Akana *et al.*, "Classroom sound can be used to classify teaching practices in college science courses," *Proceedings of the National Academy of Sciences*, vol. 114, no. 12, pp. 3085–3090, 2017.

[5] R. Cosbey, A. Wusterbarth, and B. Hutchinson, "Deep learning for classroom activity detection from audio," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 3727–3731.

[6] E. Slyman, C. Daw, M. Skrabut, A. Usenko, and B. Hutchinson, "Fine-grained classroom activity detection from audio with neural networks," *arXiv preprint arXiv:2107.14369*, 2021.

[7] "Earlyspark organization website," https://www.earlyspark.org/, [Online; last accessed 18-June-2022].

[8] L. Wan, Q. Wang, A. Papir, and I. L. Moreno, "Generalized end-to-end loss for speaker verification," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 4879–4883.

[9] E. Fonseca, M. Plakal, D. P. Ellis, F. Font, X. Favory, and X. Serra, "Learning sound event classifiers from web audio with noisy labels," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 21–25.

[10] I. D. Gaspón and I. Saratxaga, "Audio scene classification using enhanced convolutional neural networks for DCASE 2021 challenge."

[11] B. McFee, C. Raffel, D. Liang, D. P. Ellis, M. McVicar, E. Battenberg, and O. Nieto, "librosa: Audio and music signal analysis in python," in *Proceedings of the 14th python in science conference*, vol. 8, 2015, pp. 18–25.

[12] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: an ASR corpus based on public domain audio books," in *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2015, pp. 5206–5210.

[13] Z. Fan, M. Li, S. Zhou, and B. Xu, "Exploring wav2vec 2.0 on speaker verification and language identification," *arXiv preprint arXiv:2012.06185*, 2020.

[14] N. Vaessen and D. A. Van Leeuwen, "Fine-tuning wav2vec2 for speaker recognition," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 7967–7971.

[15] J. Cramer, H.-H. Wu, J. Salamon, and J. P. Bello, "Look, listen, and learn more: Design choices for deep audio embeddings," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 3852–3856.

[16] J. F. Gemmeke, D. P. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter, "Audio set: An ontology and human-labeled dataset for audio events," in *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2017, pp. 776–780.

[17] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[18] X. Xia, R. Togneri, F. Sohel, and D. Huang, "Frame-wise dynamic threshold based polyphonic acoustic event detection," 2017.

[19] J. Žibert, A. Brodnik, and F. Mihelič, "An adaptive BIC approach for robust speaker change detection in continuous audio streams," in *International Conference on Text, Speech and Dialogue*. Springer, 2009, pp. 202–209.

[20] M. Marolt, C. Bohak, A. Kavčič, and M. Pesek, "Automatic segmentation of ethnomusicological field recordings," *Applied Sciences*, vol. 9, no. 3, p. 439, 2019.

[21] F. Eyben, F. Weninger, F. Gross, and B. Schuller, "Recent developments in openSMILE, the munich open-source multimedia feature extractor," in *Proceedings of the 21st ACM international conference on Multimedia*, 2013, pp. 835–838.

[22] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, "mixup: Beyond empirical risk minimization," *arXiv preprint arXiv:1710.09412*, 2017.