Unsupervised Methods for Audio Classification from Lecture Discussion Recordings

Hang Su[1], *Borislav Dzodzo*[1], *Xixin Wu*[1], *Xunying Liu*[1], *Helen Meng*[1]

[1]Department of Systems Engineering and Engineering Management,
The Chinese University of Hong Kong, China

{hangsu,bdzodzo,wuxx,xyliu,hmmeng}@se.cuhk.edu.hk

Abstract

Time allocated for lecturing and student discussions is an important indicator of classroom quality assessment. Automated classification of lecture and discussion recording segments can serve as an indicator of classroom activity in a flipped classroom setting. Segments of lecture are primarily the speech of the lecturer, while segments of discussion include student speech, silence and noise. Multiple audio recorders simultaneously document all class activities. Recordings are coarsely synchronized to a common start time. We note that the lecturer's speech tends to be common across recordings, but student discussions are captured only in the nearby device(s). Therefore, we window each recording at 0.5 s to 5 s duration and 0.1 s analysis rate. We compute the normalized similarity between a given window and temporally proximate window segments in other recordings. Histogram plot categorizes higher similarity windows as lecture and lower ones as discussion. To improve the classification performance, high energy lecture windows and windows with very high and very low similarity are used to train a supervised model, in order to regenerate the classification results of remaining windows. Experimental results show that binary classification accuracy improves from 96.84% to 97.37%.

Index Terms: audio classification, unsupervised classification, flipped classroom

1. Introduction

In a group-based flipped classroom setting, basic knowledge is studied at home while class time is reserved for advanced lecture and in-class group discussion [1, 2, 3]. Educational research findings have shown that the successful conduct of productive student discussion is correlated with high quality learning. Peer-to-peer discussions and explanations reinforce newly acquired knowledge and is encouraged by educational guidelines [4, 5]. In-class group exercise with discussions have been shown to improve class material comprehension of both high and low student performers [6]. Previous findings [7] state that student discussion is significantly and positively related to critical thinking. Therefore, an indicator of class quality assessment can be obtained by means of detecting the amount of time provided for student discussion. Automated classification detects if the recording segments are *lecture* or *discussion* in a university group-based flipped class. Based on binary classification result, the amount of time allocated for the student discussion can be easily deduced.

In order to properly classify classroom audio into different categories, most previous classroom audio classification methods applied supervised machine learning. Owens et al. [8] developed a supervised machine learning method to distinguish between different class conduct modes such as single voice, multiple voices and no voice. Anusha et al. [9] used the LIUM speaker diarization toolkit and manually acquired ground-truth labels to automatically classify teacher's speech, children speech and their overlap with 77.3%, 71.6%, and 3.1% accuracies, respectively. A supervised, Naïve Bayes based, multiple microphone analysis classified class activity into Question and Answers, Procedures and Directions, Seatwork, Group and Lecture [10]. Although supervised methods are usually more accurate and allow classification of audio into more categories, manual labelling is a resource-intensive task. Minimal manual labelling is preferred by educators since they require rapid evaluation of classroom activity without additional workload. Moreover, some previous classroom audio classification methods relied on participants to wear *individual* microphones [11, 12]. Language ENvironment Analysis (LENA) has been used to achieve automated unsupervised classification of class activities [12], but participants had to wear microphones which is an intrusive way of audio recording.

In the current study, which aims to simplify deployment and thereby democratize Educational Data Mining, audio in the flipped classroom is captured through multiple recorders placed unobtrusively around the classroom, but not worn by any individuals, and processed for educational indicators automatically without the need for additional manual labour. The lecture audio segments are mostly recorded by all microphones, while the discussion audio segments are only recorded by the nearby microphones. The audio recordings are windowed at 0.5 s to 5 s duration at 0.1 s analysis rate. Our objective is to devise a method for reliable binary classification of windowed segments of *lecture*, against windowed segments of *discussion*. Normalized similarity is computed between a given window segment and those in other recordings that lie in temporal vicinity. An unsupervised approach based on a histogram count is used for dividing between higher similarity windows as lecture and lower similarity windows as student discussion. Then, further improvement is achieved with a boosted unsupervised framework. High confidence labeled windows contain *lecture* with high energy and very high similarity as well as *discussion* with very low similarity. The windows labeled with high confidence are used to train a neural network in supervised fashion, which is used to re-classify the remaining segments. Experiment results show that boosted unsupervised framework achieve higher classification accuracy.

The paper is organized as follows: The second section introduces the data that was recorded in a group-based, flipped classroom of a university-level engineering mathematics course. The third section illustrates the unsupervised algorithm for this classification task. The boosted unsupervised framework is described in section four. The conclusions and future work are stated in section five.

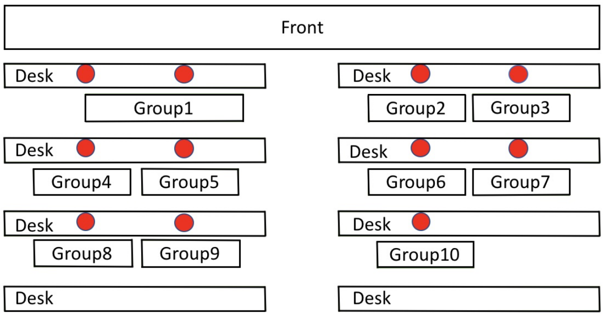

Figure 1: *Audio recording environment*

2. Flipped Classroom Audio Corpus

Data was collected from an elite flipped course of Linear Algebra at The Chinese University of Hong Kong [13]. In this course, students watched online instructional videos at home, while class time consisted of advanced lecture and problem solving in group. Three or four students formed a study group. Audio recording was conducted in every study group simultaneously in order to record their group discussion for further learning analysis and finding indicators of class assessment. The recording was started by a press of the button for each device, and the recording start time might differ by up to 10 minutes. The TASCAM DR-05 recorder was used with a 44.1 kHZ sampling rate setting. Recorders were at least 1.5 meters apart and are positioned to face the student groups, as illustrated by the red markings in Figure 1. Non-intrusive multiple microphone recording method allows the study participants to engage in their activities instead of focusing on the recording environment. After initial setup, this method of data collection allows for continuous and automated collection and analysis of classroom activities. Every audio recording contains the lecture speech by the professor or teaching assistant(s), the speech from student group discussions, noise and silence. The lecture time mostly contains lecture, while the discussion time contains discussion speech, noise and silence. In this work, the lecture speech is referred to as *lecture*, and the remaining speech is referred to as *discussion*

3. Unsupervised Classification of Lecture versus Discussion

The audio recording method for reference [12] and this study are quite different, therefore we can not directly use their audio processing method to our recordings. The distance between speaker and microphone is constant in their study, therefore energy is a reliable indicator under the circumstances of their study, whereas distance varies in this study. In designing the classification approach, we aim to fully leverage the redundancy across simultaneous recordings from the devices placed around the classroom. Specifically, lecture will have highly similar windowed segments across various recordings. Since Mel Frequency Cepstral Coefficient (MFCC) has been widely used in the literature to compute similarity between audios [14, 15, 16], we extract MFCC from all of the audios that were recorded from the same class, and one recording is picked as a timing reference. The recording is started by a press of the button for each device, and the recording start time may differ by up to 10 minutes. Therefore, it is important to synchronize the various recordings to form the search space for window comparisons.

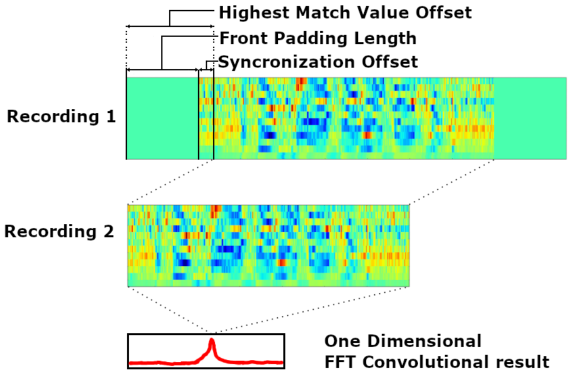

Figure 2: *Synchronization between two recordings*

Figure 2 shows how synchronization is achieved by convolution of the Fast Fourier Transform (FFT) between a selected recording with a reference recording. We note that after this procedure, an synchronization error of up to 0.5 second remains is observed.

All possible pairs of synchronized recordings are taken into further consideration. As illustrated in Figure 3 each window from the first pair member is FFT convoluted with 10 s search space of closest aligned time in the second pair member. A match is located as the point of highest value from convolution and it is recorded in the match results array for each window in the first pair member. Each offset between two best matching windows is also recorded for each window in the first pair member. All match result arrays are normalized to a range between 0 and 1000. All match results that represent the same recording time are averaged in order to get an average match result for each window in the reference recording. Then, the algorithm creates a histogram of averaged match results shown in Figure 4. The histogram is a bimodal distribution. The peak located close to match value 200 indicates the mode of *discussion* window match values, while the peak located close to match value 700 indicates the mode of *lecture* window match values. Therefore, by searching for the lowest y-axis value between the two peaks in the bimodal distribution, the threshold for classifying *lecture* and *discussion* can be selected automatically. Windows that have final matches greater than the threshold are classified as *lecture* and lower than the threshold are classified as *discussion*. Although classification results are only calculated for the reference recording, other recordings share the same classification result for the same moment of time. Classification results for other recordings are created by time shifting the reference recording. Time shift is the mode of offsets obtained in every 10 s search between reference and target recording.

Table 1: *Classification Accuracy for Different Window Sizes*

Window Size (sec)	Best threshold Accuracy	Bottom threshold Accuracy
0.5	94.77%	93.28%
0.9	96.61%	96.34%
2	**96.90%**	**96.84%**
3	96.61%	96.43%
4	96.40%	96.16%
5	96.29%	96.07%

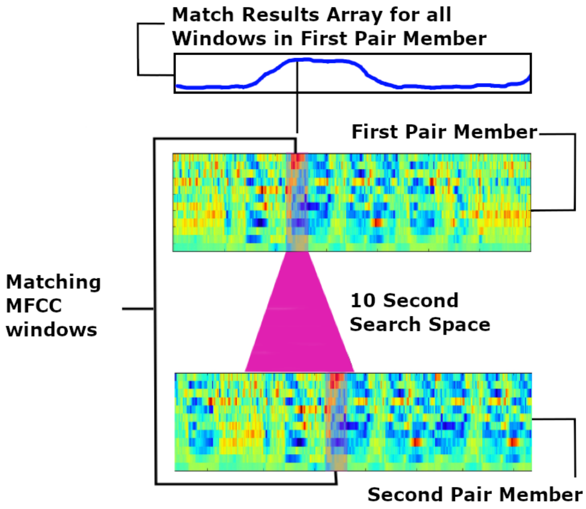

Figure 3: *MFCC Window matching*

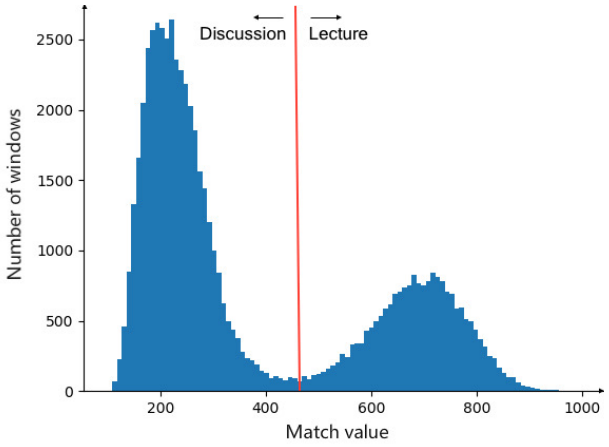

Figure 4: *Histogram of match values to a reference recording. The decision threshold selected is shown by the red line*

3.1. Result of Unsupervised Binary Classification

3.1.1. Experimental Setup

Five recordings from different classes are randomly selected and manually labeled. The entire audio is labeled as either *lecture* or *discussion* with 0.1 s resolution as defined in section2. If two lecturing segments from the same recording are less than 1 s apart, then the intermediate time is also labelled as *lecture*. Overlap of lecturing speech and student conversation is labelled as *lecture*. Annotation labels are created and double checked by two people. These labelled files are only used for testing the performance of our proposed methods.

Since our algorithm is designed to simultaneously process all of the recordings for one class, the algorithm is evaluated in terms of classification accuracy on recordings taken from five different classes. For every 0.1 s of recording time, classification results of the algorithm are compared against manually annotated labels. Accuracy is calculated according to Equation (1)

$$ACC = \frac{N_{common}}{N_{all}} \tag{1}$$

where N_{common} represents the number of matching annotation and classification labels, and N_{all} represents the total number of labels in all five test files.

Table 2: *Confusion matrix of 2 s window classification result*

		Prediction		
		Discussion	Lecture	Recall
Actual	Discussion	173313	2464	98.60%
	Lecture	5284	64178	92.39%
	Precision	97.04%	96.30%	

3.1.2. Results and Discussion

Window size is varied in the experiment. It can be observed in Table 1 that the 2 s window performs better than other window sizes in terms of classifying with both the bottom threshold and the best threshold. The bottom threshold is selected by the proposed algorithm. Every threshold value between 0 and 1000 is evaluated in terms of classification accuracy against annotated labels in order to find the best threshold. The accuracy of bottom threshold is nearly the same as the best threshold; therefore, the search for the best threshold can be avoided by the use of the bottom threshold which can be automatically selected without annotation labels.

The confusion matrix of 2 s window is shown in Table 2. Occasionally, student speech overlaps with lecturing speech, which makes the averaged match value of the *lecture* lower than the threshold and classified as *discussion*. Window match values that cause this kind of error are usually marginally lower than the threshold, because only some of the recordings contain an overlap. In addition, some noise and silence have high averaged match value across all of the recordings, so they are incorrectly classified as *lecture*. Furthermore, if one or two groups talk loudly, while other groups are nearly silent or talk quietly, then some pairs of recorders have high similarity with each other, and the averaged match value may be marginally higher than the threshold, which leads to incorrect classification as *lecture*. It is observed that there are 193163 windows that have match values higher than $(threshold + 150)$ and simultaneously have energy higher than half of the whole recording, or have a similarity lower than $(threshold - 150)$. Among these 193163 windows there are 741 windows that are incorrectly classified. This indicates that although almost 20% of the windows don't satisfy these conditions, condition satisfying windows are classified correctly with 99.62% precision.

4. Boosted Unsupervised Framework

4.1. Overview

Section 3.1.2 discussed the error cases and found that the error is very low in windows classified as *discussion* that have very low match values and the windows classified as *lecture* that have very high match value and high energy. Therefore, windows classified as *lecture* with high energy and very high match values are defined as high confidence labeled windows. Windows classified as *discussion* with very low match values are also defined as high confidence labeled windows. In order to enhance the unsupervised binary classification performance, we assume that the windows labeled as *lecture* with high confidence are ”true” *lecture* windows, and other windows labeled as *discussion* with high confidence are ”true” *discussion* windows. These are then used as labels to train a supervised classifier,

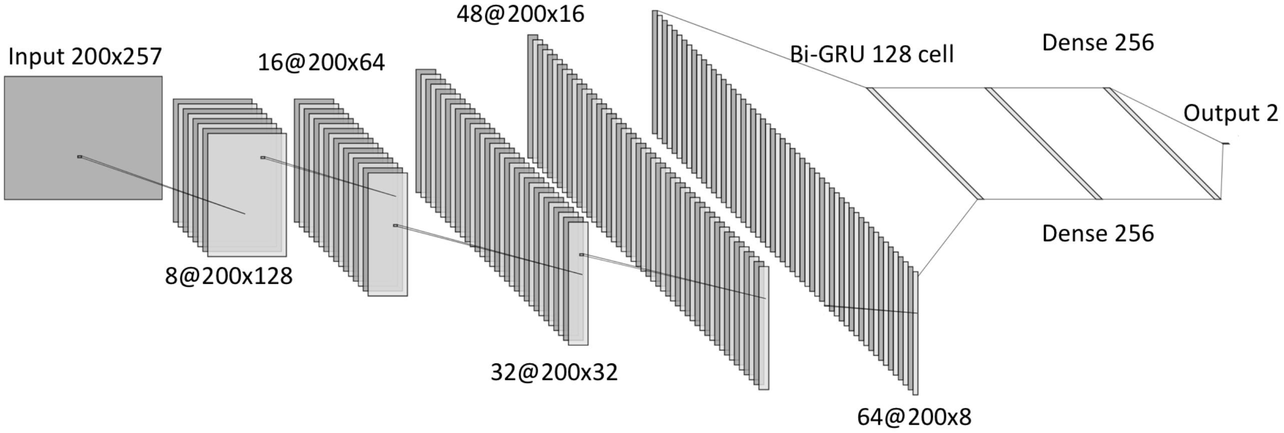

Figure 5: *Neural Network Architecture*

which re-classifies the remaining windows.

4.2. Experiment

4.2.1. Data Pre-processing for Supervised Learning

Since some previous audio processing works have shown the benefit of directly applying CNNs onto spectrograms [17, 18], we use spectrograms as input for our neural network with CNNs in the first five layers. Spectrograms are calculated by Short-time Fourier transform with 10-ms Hanning window, 10-ms shift and 512-point FFT for each recording file. High confidence windows with 2 s window size setting are selected as training and validation data. The *discussion* windows that have match values less than $(threshold - 150)$ are selected as *discussion* training and validation data. The *lecture* windows that exceed energy levels greater than half of the average energy of the whole recording and simultaneously have match value larger than $(threshold + 150)$ are selected as *lecture* training and validation data. The remaining windows are reclassified after the model is trained by high confidence training data. The ratio of training set size and validation set size is $10:1$. The validation set is only used for selecting the model.

4.2.2. Neural Network Architecture

The supervised learning model is implemented as a neural network architecture that is demonstrated in figure5. The input of the neural network is a spectrogram, represented by a 200×257 matrix, where 200 represents the time domain and 257 is the frequency domain. Spectrogram is first processed by 5 CNN layers, in order to extract the higher level features from time and frequency information. Then, a Gated Recurrent Unit (GRU) layer with 128 cells connect with the CNN layers, in order to capture the temporal information from the spectrogram. The GRU input shape is $200 \times 8 \times 64$, representing timesteps$\times$frequency dimensions$\times$output channels. Only the last timestep output of GRU are considered, with 256 dimensions. Finally, three dense layers are connected to the GRU with 256, 256, 2 nodes, in order to perform the classification.

4.2.3. Experiment Results and Discussion

The experiment is evaluated on the five manually labeled files, in order to compare against the initial unsupervised algorithm. The experiment is conducted 5 times and the median result shows that accuracy of the boosted unsupervised framework as 97.37%, which outperforms 96.84% for the initial 2 s window result shown in Table 1. Since the high confidence windows result from the boosted unsupervised framework are the same as the initial unsupervised algorithm, the improvement entirely comes from the reclassification of the remaining windows. The confusion matrix of the boosted unsupervised framework is shown in Table 3. Compared with the Table 2, the number of both lecture and discussion error windows is lower, which demonstrates the superior performance of the boosted unsupervised framework.

Table 3: *Confusion matrix of boosted unsupervised framework*

		Prediction		
		Discussion	Lecture	Recall
Actual	Discussion	173592	2185	98.76%
	Lecture	4259	65203	93.87%
	Precision	97.61%	96.76%	

5. Conclusion and Future Work

In this work multiple standalone recorders were utilized in a non-intrusive audio analysis study of a flipped classroom in order to assess the amount of time allocated to lecture and group discussion, which is a documented educational assessment metric. Synchronization of class recordings with an FFT convolution was conducted in order to compare temporary close windows of time in all of the class recordings and reduce the analysis window from 10 minutes to 10 seconds. Unsupervised algorithm achieved a high classification accuracy against manually annotated labels. Then, a boosted unsupervised framework was devised and outperformed the unsupervised algorithm, by using high confidence classification results to train a supervised neural network. In the future, we plan to develop an unsupervised method to further classify the silence and noise in our *discussion* category and apply the results to further educational data mining.

6. Acknowledgements

This paper is partially supported by the CUHK TDLEG Grant (2016-2019) and a grant from the Dr. Stanley Ho Medical Development Foundation.

7. References

[1] J. W. Baker, "The" classroom flip," *Using web course management tools to become the guide by the side*, 2000.

[2] K. Ash, "Educators evaluate flipped classrooms," *Education Week*, vol. 32, no. 2, pp. s6–s8, 2012.

[3] J. Bergmann and A. Sams, *Flip your classroom: Reach every student in every class every day*. International society for technology in education, 2012.

[4] N. C. of Teachers of Mathematics. Commission on Teaching Standards for School Mathematics, *Professional standards for teaching mathematics*. Natl Council of Teachers of, 1991.

[5] N. C. of Teachers of Mathematics, *Principles and standards for school mathematics*. National Council of Teachers of, 2000, vol. 1.

[6] M. K. Smith, W. B. Wood, K. Krauter, and J. K. Knight, "Combining peer discussion with instructor explanation increases student learning from in-class concept questions," *CBELife Sciences Education*, vol. 10, no. 1, pp. 55–63, 2011.

[7] D. G. Smith, "College classroom interactions and critical thinking." *Journal of Educational Psychology*, vol. 69, no. 2, p. 180, 1977.

[8] 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.

[9] A. James, V. Y. Chua, T. Maszczyk, A. M. Nunez, R. Bull, K. Lee, and J. Dauwels, "Automated classification of classroom climate by audio analysis," in *International Workshop on Spoken Dialog System Technology*, 2018.

[10] P. J. Donnelly, N. Blanchard, B. Samei, A. M. Olney, X. Sun, B. Ward, S. Kelly, M. Nystrand, and S. K. D'Mello, "Multi-sensor modeling of teacher instructional segments in live classrooms," in *Proceedings of the 18th ACM international conference on multimodal interaction*. ACM, 2016, pp. 177–184.

[11] 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.

[12] Z. Wang, K. Miller, and K. Cortina, *Using the LENA in Teacher Training: Promoting Student Involement through automated feedback*. na, 2013, vol. 4.

[13] S. Jaggi, X. Wang, B. Dzodzo, Y. Jiang, and H. Meng, "Systematic and quantifiable approach to teaching elite students," 2018, eurasian Conference on Educational Innovation. [Online]. Available: https://tinyurl.com/yyrfrqtk

[14] L. Muda, M. Begam, and I. Elamvazuthi, "Voice recognition algorithms using mel frequency cepstral coefficient (mfcc) and dynamic time warping (dtw) techniques," *arXiv preprint arXiv:1003.4083*, 2010.

[15] S. D. Dhingra, G. Nijhawan, and P. Pandit, "Isolated speech recognition using mfcc and dtw," *International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering*, vol. 2, no. 8, pp. 4085–4092, 2013.

[16] T. Muttaqi, S. H. Mousavinezhad, and S. Mahamud, "User identification system using biometrics speaker recognition by mfcc and dtw along with signal processing package," in *2018 IEEE International Conference on Electro/Information Technology (EIT)*, May 2018, pp. 0079–0083.

[17] A. Satt, S. Rozenberg, and R. Hoory, "Efficient emotion recognition from speech using deep learning on spectrograms." in *INTERSPEECH*, 2017, pp. 1089–1093.

[18] Y. M. Costa, L. S. Oliveira, and C. N. Silla Jr, "An evaluation of convolutional neural networks for music classification using spectrograms," *Applied soft computing*, vol. 52, pp. 28–38, 2017.