Speaker Diarization in the Classroom: How Much Does Each Student Speak in Group Discussions?

Jiani Wang
Worcester Polytechnic Institute
jwang21@wpi.edu

Shiran Dudy
Northeastern University
shirdu2@gmail.com

Xinlu He
Worcester Polytechnic Institute
xhe4@wpi.edu

Zhiyong Wang
University of Colorado Boulder
zhiyong.wang@colorado.edu

Rosy Southwell
University of Colorado Boulder
roso8920@colorado.edu

Jacob Whitehill
Worcester Polytechnic Institute
jrwhitehill@wpi.edu

ABSTRACT

One important dimension of classroom group dynamics & collaboration is how much each person contributes to the discussion. With the goal of measuring how much each student speaks, we investigate how automatic speaker diarization can be built to handle real-world classroom group discussions. We examine key design considerations such as the level of granularity of speaker assignment, speech enhancement techniques, voice activity detection, and embedding assignment method, so as to find an effective configuration. The best speaker diarization that we found was based on the ECAPA-TDNN speaker embedding model and used Whisper automatic speech recognition to find speech segments. Diarization error rates (DER) on challenging noisy spontaneous classroom data were around 34%, and the correlations of estimated vs. human annotations of how much each student spoke reached 0.62. The presented diarization system has potential to benefit educational research and also to give teachers and students useful feedback to understand their group dynamics.

Keywords

Speaker Diarization, automatic speech recognition, automatic classroom analysis, group collaboration

1. INTRODUCTION

In modern classroom learning, it is vital that students not only learn academic subject-matter such as math, reading, and writing, but also that they learn broader critical-thinking, communication, and collaborative learning skills [7]. These skills are not confined to any particular subject but permeate students' entire learning journey. Instrumental to nurturing these skills is to incorporate discussions into the classroom, either in the classroom as a whole or in small groups. Students who participate in these discussions actively tend to achieve better learning outcomes than those who do not [15], and there is a strong relationship between the frequency and quality of student talk during a lesson and student achievement [27]. Asking, explaining and discussing with others helps to stimulate students' thinking and to deepen their memory of the curricula. Thus, the amount of "talking time" by each student is an important metric with which to analyze a student's learning.

Measuring classroom speech: Due to the importance of fostering effective student collaboration in group discussions, it could be beneficial to educators to gauge automatically students' behaviors in classroom group discussions, both to facilitate large-scale research studies and to provide learners with feedback. Given the complexity of real-world classroom dynamics, measuring classroom speech is very challenging. Traditional methods of assessing classroom activities, such as inviting experts into the classroom and recording discussions manually, or relying on survey questionnaires, cannot provide a comprehensive assessment of students' performance in group discussions [22]. Such methods are labor-intensive, time-consuming, prone to subjective biases, and thus may lack accuracy and objectivity. Therefore, having an automatic tool that can assist teacher in understanding the discussions of each group and the performance of students within each group would be very useful.

Speaker diarization for classroom discussion analysis: Modern deep learning-based speech analysis algorithms offer new ways to measure the quality and quantity of classroom speech, and they can avoid some problems and risks associated with traditional assessment. Speaker diarization algorithms, in particular, identify "who is speaking when" automatically [2]. Speaker diarization can help educators to understand students' degree of participation and to assess their communication and collaboration skills. To date, however, there is a lack of research on how speaker diarization can be deployed in real-world, noisy classrooms with unscripted speech of children. Our paper seeks to help fill this gap. In contrast to other classroom analytic methods that assume specialized hardware (e.g., one LENA microphone for each child), our work requires only a single table-top microphone for each group, thus being easier to deploy and less obtrusive.

Research contribution: This paper explores systematically how to design and implement a robust and accurate speaker diarization system which is capable of identifying speech segments and corresponding speakers during classroom group discussions. Due to privacy concerns, we focus on locally

deployable solutions rather than cloud-based diarization services. We present a general Speaker Diarization Framework and describe how it can be deployed in real-world classrooms to quantify how much each person contributes to a group discussion. The significance of this research lies in its potential to drive changes in educators' classroom management and student assessment methods. Firstly, the automated speech recognition and speaker identification processes alleviate teachers' burdens in classroom supervision, thus enabling them to devote more time and energy to fostering meaningful discussions and interactions with students. Secondly, the objective data collected through automatic assessment methods provide educators with a more comprehensive and objective basis for evaluating student performance, promoting fairness and effectiveness in education practices. Furthermore, each student is a unique individual, and through the detailed feedback provided by the system for each student, educators can provide personalized guidance to them.

Figure 1: Classroom setup of our study, containing multiple groups of interacting students.

2. RELATED WORK

2.1 Speech Analysis of Classroom Interactions

In educational data mining and learning analytics, there are numerous applications of speech processing methods. Beccaro *et al.* utilized speaker diarization as the core method to build a speech processing model for assessing students' performance and engagement during oral exams [3]. They then examined the correlation between the emotional expressions of students during speech and their final oral examination scores. Gomez *et al.* also employed speaker diarization for classroom analysis which is the similar application as ours [11]. They confronted the same challenges as us of limited data with substantial noise. Instead of using deep learning methods, they solved the problem by physical principles and used virtual microphones. They computed the spatial information of the speakers based on the speaker geometry and estimated the room impulse responses (RIRs). Ultimately, they got the predicted speakers according to the cross-correlation matrix calculated based on RIRs. Olney *et al.* proposed a method to deal with the class imbalance problem [19]. Cao *et al.* investigated the impact of ASR errors on the analysis of collaborative class and provided constructive suggestions on optimizing group discourse modeling tasks [5]. Dutta *et al.* proposed a translation framework which applied automatic speech recognition (ASR) to track preschool children's conversational speech [9]. Kelly *et al.* applied ASR to detect authentic questions in classroom in order to support teaching effectiveness improvement [16].

2.2 Speaker Diarization

Speaker Diarization aims to automatically identify "who speaks when" within an input audio [20]. There are various mature methodologies to achieve this, including feature embeddings [26], speaker modeling [25] [18], segmentation and clustering algorithms [17] as well as end-to-end methods [31, 10, 14]. In recent years, an increasing number of approaches based on deep learning models have been proposed for speaker diarization. Desplanques *et al.* used emphasized channel attention, propagation and aggregation deep learning model based on the Time-Delay Neural Network (TDNN) based system, named ECAPA-TDNN [8]. In this model, they applied architectural enhancements, additional skip connections and channel attention to improve the performance. Chen *et al.* proposed WavLM[6] to solve full-stack downstream speech tasks. It employs gated relative position bias for the Transformer structure and jointly learns masked speech prediction and denoising in pre-training. WavLM achieves SOTA performance on the CALLHOME speaker diarization benchmark. Finally, Amazon [1], Google [12] and other companies offer cloud-based diarization services, but for many schools these services are unacceptable due to privacy concerns.

3. DATASET

In our study we used the Sensor Immersion dataset [29], including both the enrollment and test audios. Sensor Immersion was collected "in-the-wild" from middle- and high school classrooms in the western United States (see Figure 1). It consists of 32 audio recordings, each of which is approximately 5 minutes long, all of which were unscripted and contain authentic student interactions. Each audio was recorded of a group discussion of 2 to 4 students who are discussing how to use different sensors (temperature, moisture, CO_2, etc.) to complete a collaborative science task. Rather than give each student their own microphone, which is inconvenient and arguably intrusive for both teachers and students, we used omnidirectional table-top microphones to record the audios. Due to the presence of multiple discussion groups within the same classroom simultaneously, the audios contain significant environmental noise. The proportions of audio containing different numbers of simultaneous speakers are shown in Table 1.

# Speakers	0	1	2	3
Proportion	63.37%	35.02%	1.60%	0.01%

Table 1: Proportion of simultaneous speech from different numbers of speakers.

3.1 Speaker Enrollment

Prior to each group discussion, the students in the group "enroll" themselves by recording a sentence of their voice and giving their name. Each enrollment (at least 5 sec long)

is recorded by a student in each group and typically encompasses a short greeting and the student's name. The goal is for each student to give a clean and short (5 sec) recording of only their speech so that the diarization system can learn what their voice sounds like. These enrollment audios are not a part of the classroom group discussion themselves but are recorded beforehand. For each enrollment, there is only one speaker, thus avoiding the case that multiple speakers are talking simultaneously. However, it often still contains background noise. Each student possesses only one enrollment recording. The teachers and the other researchers in the classroom do not have enrollments and thus we treat their speech as background noise. redGiven teacher enrollments, however, it would be straightforward to detect speaker speech.

3.2 Annotation

All of the audios in our dataset were manually labeled for who-spoke-when. In particular, each utterance that was spoken by a student was annotated for the start and end times, as well as the content of what was said. These labels enable us to analyze how accurately an automatic speaker diarization system can perform on the dataset.

3.3 Challenges

In the Sensor Immersion setting, students are divided into groups of 2-4 people, where each group is recorded by a table-top omnidirectional microphone. However, as all groups are in the same classroom, each microphone not only captures the voice of its own group but also records the voice of the others. Also, due to the scarcity of actual classroom data (only 32 recordings), which are reserved for testing the system's efficacy, we are unable to use this data for training or fine-tuning models.

4. SPEAKER DIARIZATION FRAMEWORK

Here we describe the general framework we used to perform speaker diarization, including the different design variants that we explored. The inputs to our speaker diarization system are always (1) a single short audio "enrollment" clip (e.g., "Hi, my name is [name] and this is my voice.") from each student and (2) a test audio that the user wishes to diarize. Our diarization system then proceeds in several phases (see Figure 2), described in the next subsections.

4.1 Speech Enhancement

Unconstrained audiorecordings from table-top microphones in school classrooms involving multiple simultaneous discussions from different groups of students can be highly noisy. Hence, as an optional initial step, we can try to improve the speech quality by filtering it with a speech enhancement system, the Speechbrain Waveform transform enhancer [24].

For the enrollments, we apply the enhancer directly on the original enrollment audios. For the test audios, we apply it to the active speech segments obtained from the Voice Activity Detector (described in the next section).

4.2 Voice Activity Detector

Common to many automatic speech applications is the use of a Voice Activity Detector (VAD) to find segments of the entire input audio that contain any speech at all. All segments containing no speech according to the VAD forgo any further processing and are immediately classified as "no speech" in our diarization system. For sure, there will also be errors where segments with speech are incorrectly identified as "no speech". These error are reflected in the DER calculation as a Missed Detection.

We apply the VAD to the enrollments and the test audios in different ways, as described below.

4.2.1 Enrollments

The enrollment audios in our study were often very noisy and also contained segments without any speech. In an effort to improve the quality of the enrollment audios, we thus explored applying a VAD to select the most useful portions of them. In particular, we used the SpeechBrain CRDNN VAD [24] to find the speech segments with the highest probability of speech.

4.2.2 Test audio

For processing the test audios, we explored three different ways of detecting moments of non-speech: Whisper ASR, non-speech enrollments, and a secondary VAD system (either SpeechBrain CRDNN or Silero [30]). We also tried combinations of these approaches, as described below.

Whisper: We found in pilot experiments that the Whisper [23] automatic speech recognition system can effectively be used as a VAD. Whisper produces for any given input audio a list of starting and ending timestamps of spoken sentences, along with the estimated transcript of what was said, and these timestamps largely agree with the periods of speech during the entire test audio. For our application, we ignore the transcript and use just the timestamps. Using Whisper is often also convenient for various downstream applications (e.g., inferring who said what in a discussion group).

Non-speech enrollments: As an additional way to detect non-speech moments on the test audios, we tried comparing the extracted speech embeddings to an embedded audio of background noise. For each audio, we extracted a segment ($\geq$ 5 seconds) which has no one speaking and only contains background noise. This segment serves as the "non-speech enrollment". Subsequently, this enrollment is treated on par with the enrollments of candidate speakers. Hence, when using this technique, the enrollment set for each test audio comprises enrollments from all speakers present as well as the non-speech enrollment.

Secondary VAD: We tried combining Whisper ASR as a first-stage VAD (either SpeechBrain CRDNN or Silero). When utilizing two VAD models concurrently, for each test audio, we initially apply Whisper for the first round of voice activity detection. This yields intervals containing at least one speaker. Subsequently, based on the start and end times of the intervals, we extract corresponding segments from the test audios, named Whisper-segments, which serve as the inputs for the next VAD model (either SpeechBrain CRDNN or Silero). In the second round, the second VAD model takes the Whisper-segments as inputs and computes the probability of speech for each segment, named $pred_{speech}$. If $pred_{speech}$ exceeds a predefined threshold, the segment is identified as containing at least one speaker and its corre-

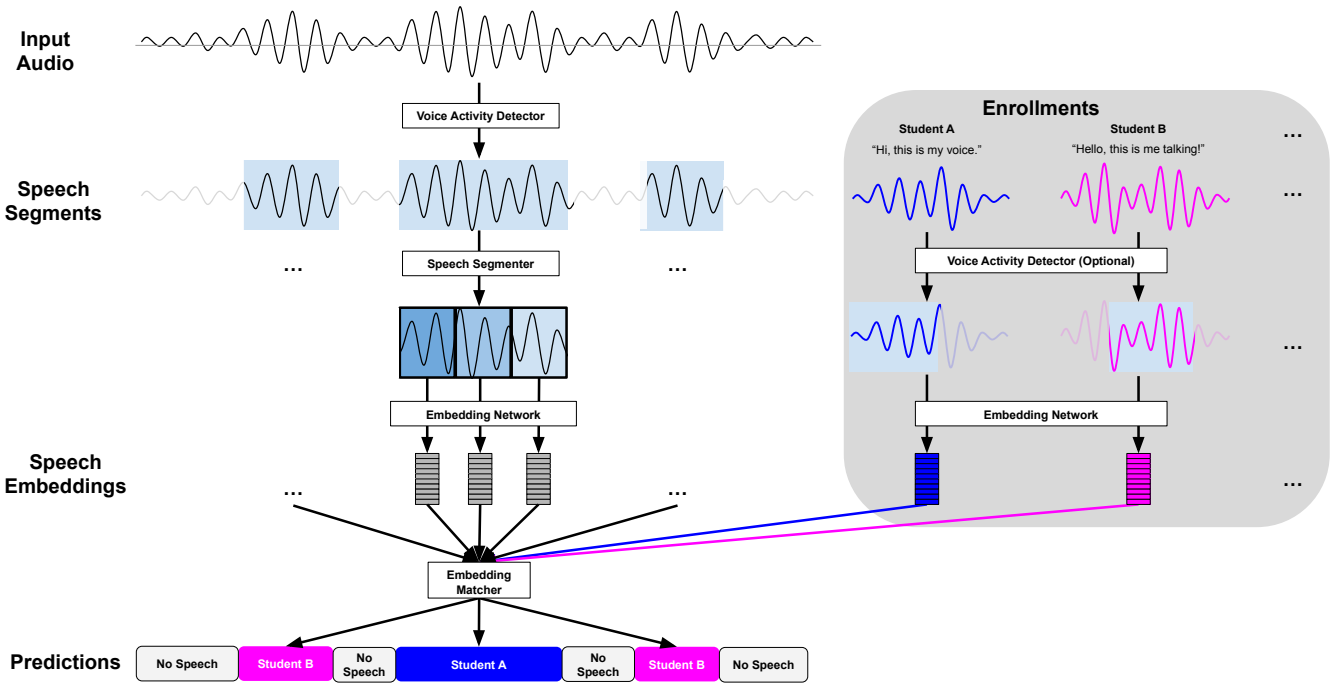

Figure 2: Speaker Diarization Framework

sponding intervals will be saved for later compute its embedding. If $pred_{speech}$ falls below the threshold, the segment is immediately classified as "no speech".

VAD threshold: we employed the threshold of 0.8798 for SpeechBrain CRDNN for most of the experiments and we selected this value by the following processes. We calculated the averaged speech probability for the segments with at least one speaker, named $prob_{act}$, and the averaged speech probability for the segments without any speaker, named $prob_{de}$, and finally got the average of $prob_{act}$ and $prob_{de}$ as the threshold of the CRDNN.

Combination of VAD model and non-speech enrollments: When employing both VAD model(s) and non-speech enrollments, we first utilize VAD model(s) to get the intervals that contain at least one speaker. Subsequently, the result intervals will be used to extract the corresponding segments and then obtain embeddings. For each embedding, in addition to computing cosine similarity with the embeddings of all candidate speakers' enrollments, it is also compared against the embedding of a non-speech enrollment. If the cosine similarity with the non-speech enrollment embedding is the highest, the corresponding segment is labeled as "no speech".

4.3 Speech Segmenter

Given the sentences of detected speech (represented in blue in Figure 2), we may either split them up further into fixed-length frames, or process each one as a whole. When splitting into frames, we used a width of 2 seconds and a timestep of 0.75 seconds.

4.4 Embedding Network

The essence of any speaker diarization system is a function that maps a segment of speech into an embedding space such that embeddings from the same speaker are close together and embeddings from different speakers are far apart. As the model architecture, we use the Emphasized Channel Attention, Propagation and Aggregation in Time Delay Neural Network (ECAPA-TDNN) model [8], which (as of 2024) is state-of-the-art. We use either the off-the-shelf (pre-trained) or a fine-tuned model (see below) to extract embeddings. ECAPA-TDNN can take a variable-length audio segment as input, and thus can be applied to either an entire Whisper segment or an individual frame extracted from within a Whisper segment. In addition to processing each segment of the test audio, we also compute embeddings for each of the enrollment audio clips. Note that, in addition to ECAPA-TDNN, we also tried using WavLM; however, our pilot experiments with this model suggested it performed worse and hence we abandoned the approach.

For fine-tuning ECAPA-TDNN, we used a variety of public datasets containing children's speech, specifically the CUKids [13] (118 hours from 1354 speakers), CSLUKids [28] (98 hours from 1118 speakers), and MyST [21] (435 hours from 1300 speakers) datasets. Jointly, these datasets comprise students from ages 5-16 years and contain both scripted and spontaneous speech. We fine-tuned the off-the-shelf ECAPA-TDNN for 10 epochs at a learning rate of 0.0001.

4.5 Embedding Matcher

Given the embeddings extracted from each processed segment of input audio, as well as the embeddings of each of the enrollment audios, the next step is to assign the embeddings to speakers. In the group discussions in our dataset, it sometimes occurs that multiple students are speaking si-

multaneously. However, since the occurrence of such events was rare ($< 2\%$ of the time), we ignored such possibilities and always assigned a speech segment to a single speaker. There are two design questions for this process: whether to assign embeddings to speakers using a "nearest enrollment" vs. clustering; and the level of granularity at which the assignment is made.

4.5.1 Nearest Enrollment vs. Clustering

The usual approach to speaker diarization is to compute a cosine similarity score between the embedding of each test segment and each enrollment embedding, and then to assign the segment to the speaker with the highest similarity score. Alternatively, since we are diarizing the entire input audio offline, we can use a clustering approach: if we have k different speakers during enrollment, then we use a clustering algorithm such as k-means to cluster the test segment embeddings into k clusters. This approach may benefit from harnessing the entire trajectory of embeddings across the whole input audio. In our system, we employ the k-means algorithm, where k is set to equal to the number of speaker enrollments. The centroid of each cluster is initialized to a unique enrollment embedding. After clustering, the Hungarian algorithm is used to find the optimal matching between the clusters and the enrollments. The embeddings within the same cluster are assigned the same label.

4.5.2 Granularity of Assignment

The level of granularity at which we assign segments to speakers depends on the speech segmentation method that was used (Section 4.3). If we originally split the audio into frames, then we can either assign each frame to a speaker, or (if using Whisper as the VAD) we can aggregate the frames within each sentence to assign a single speaker to each sentence using one of several possible voting mechanisms. Alternatively, we can compute an embedding from each entire sentence as inferred by Whisper, and then directly assign each sentence to a speaker. Embedding each frame has the possible advantage that it contains "purer" segments of speech (since the probability of containing speech from multiple speakers is reduced). On the other hand, having access to longer speech segments and benefitting from the linguistic structure (which is available to Whisper since it performs speech recognition) is a possible benefit of analyzing each sentence as a whole.

For the former case, we used one of three alternative voting mechanisms as described below.

Majority Vote: Given a sentence of speech comprising frames $f_1, \ldots, f_n$, the embedding model is used to compute the embeddings $e_1, \ldots, e_n$. Then, the cosine similarity is computed between each embedding i and the enrollment embedding of each candidate speaker; the speaker with the highest cosine similarity s_i is selected as the predicted speaker p_i. Across all n frames within a sentence, the occurrences of each speaker in the predictions are tallied, and the speaker with the highest number of occurrences is chosen as the final predicted result for the sentence.

When applying the Majority Vote method, we observed instances where certain frames had the same prediction, but their corresponding cosine similarity with candidate speakers were much different. To harness the cosine similarities more fully, we thus devised the following two methods.

Weighted Vote: This method utilizes all cosine similarity values $s_1, \ldots, s_n$. After obtaining the cosine similarity between each frame and each candidate speaker, we calculate the sum of cosine similarities for each candidate speaker across all frames. The speaker with the maximum sum is then chosen as the prediction for the entire sentence. The weight of each frame is the product of the cosine similarity and the length of the frame. Thus, each cosine similarity value can make different contributions for the final prediction.

Argmax Vote: This method places emphasis on the maximum value of cosine similarity. After obtaining the cosine similarity between each frame and each candidate speaker, we select the speaker corresponding to p_{i^*} where $i^* = \arg\max_i s_i$.

5. EXPERIMENT & RESULTS

We conduct experiments on different design configurations of the Speaker Diarization Framework presented in Section 4 in order to determine which works the best.

5.1 Evaluation Metric

We measure accuracy using Diarization Error Rate (DER)[4], which measures the fraction of the total audio length in which the *set* of speakers (since there can be multiple people speaking simultaneously) was incorrectly inferred by the model. We compute this as:

$$\text{DER} = \frac{\text{False Alarm} + \text{Missed Detection} + \text{Confusion}}{\text{Total}}$$

where False Alarm is the length of speech in which no one was speaking but the model believed someone was, Missed Detection is the opposite, and Confusion means that the inferred set of people was incorrect. Note that the inferred set must exactly match the ground-truth set; otherwise, it is marked as a Confusion. For example, if during a particular moment speakers A and B were talking, but the model believed that only speaker A was talking, then this segment is considered a Confusion. Since our Speaker Diarization Framework always assigns speech segments to single speakers, it will always be penalized in the DER on any segment whose ground-truth label contains multiple speakers.

To compute the DER over our entire dataset, we compute it for each test audio individually. Then, the weighted average of the DERs over all individual audios is computed, where the weight corresponds to the length of each test audio. For significance testing between different diarization methods, we use paired t-tests across the 32 test audios.

5.2 Frame-wise vs. Sentence-wise Prediction

We compared frame-wise assignment to the four sentence-wise (sentence embedding, Majority Vote, Argmax Vote, and Weighted Vote) assignment methods. **Configuration**: In this experiment, we utilize Whisper, SpeechBrain CRDNN VAD, as well as "non-speech" enrollments as the VAD methods. We used the pre-trained ECAPA-TDNN as the embedding network and used nearest enrollment as the embedding matcher. Finally, we did not apply speech enhancement.

Hyperparameter selection: For the frame-wise approach, there are several hyperparameters that need to be chosen. We employed 5-fold cross validation to select the hyperparameters, which consisted of the *windowSize* $\in$ [0.1, 0.25, 0.5, 1, 2, 3], and *stepSize* $\in$ [0.25, 0.5, 0.75, 1] (i.e., 24 pairs of hyperparameters in total). The selection process was as follows: Initially, we evenly divided the 32 test audios into 5 groups to conduct 5 sub-experiments. For each sub-experiment, we applied the diarization pipeline to 4 groups of test audios using 24 pairs of hyperparameters and obtained 24 DER results. We chose the pair of hyperparameters which gave the lowest DER result and then measured the DER of this hyperparameter on the remaining group of test audios. Upon completing the 5 sub-experiments sequentially, we obtained 5 pairs of best hyperparameters, and we found that all 5 pairs were the same, namely: *windowSize*=2 and *stepSize*=0.75. Consequently, we selected this pair of hyperparameters for all frame-wise based experiments.

Results: The frame-wise predictions and the three voting-based methods all tied with a DER of 0.3838. Sentence-wise embedding attained a DER of 0.3937, but the difference was not stat. sig. (p =0.0748). As described in Section 4.5.2, the Majority Vote, Argmax Vote and Weighted Vote methods are based on frame-wise prediction, i.e., embeddings are extracted at the frame level, and then voting is used to aggregate at the sentence level. Since they have the same DER as frame-wise prediction, we deduce no clear benefit of the voting methods.

Due to the small accuracy difference and relative simplicity, we choose to use the sentence-embedding method for all subsequent experiments.

5.3 Pre-trained vs. Fine-tuned ECAPA-TDNN

To assess whether fine-tuning the embedding model on children's speech improved accuracy, we compared the pre-trained ECAPA-TDNN to the fine-tuned version (see Section 4.4) in terms of DER. **Configuration**: without speech enhancement; whole enrollments; Whisper, Speechbrain CRDNN, and non-speech enrollments for VAD; sentence embedding; and nearest enrollment. **Results**: The fine-tuned model achieved a DER of 0.3577, which is stat. sig. (p =0.0007) better than the pre-trained one, whose DER is 0.3937.

5.4 Speech Enhancement

We compared DER obtained with vs. without applying speech enhancement, as described in Section 4.1. **Configuration**: pre-trained ECAPA-TDNN; whole enrollments; Whisper, Speechbrain CRDNN, and non-speech enrollments for VAD; sentence embedding; and nearest enrollment. **Results**: The model without speech enhancement achieved a DER of 0.3937, which is stat. sig. (p =0.0319) better than the model with speech enhancement, whose DER is 0.4262.

5.5 Subselecting Enrollment Audio with VAD

We compared (a) using the whole enrollment audio to compute the enrollment embedding for each speaker; and (b) using only a fixed-length portion of each enrollment audio. The intuition is that we might obtain a higher-quality embedding by computing it only on the "best" parts of the enrollment audio. In particular, we selected the "best" fixed-length (for lengths 2, 4, 8, 16, 32 sec) segment within each enrollment audio according to the probability of speech output by the Speechbrain CRDNN. We then extracted an enrollment embedding from only this portion of the enrollment audio. **Configuration**: pre-trained ECAPA-TDNN; without speech enhancement; Whisper, Speechbrain CRDNN, and non-speech enrollments for VAD; sentence embedding; and nearest enrollment. **Results**: Using 4-second segmented enrollments achieved a DER of 0.3745 which is stat. sig. (p =0.0073) better than the model that used whole enrollments, whose DER is 0.3937.

5.6 Nearest Enrollment vs. Clustering

In this experiment, we compare the effectiveness of the nearest enrollment method and the clustering method (using k-means clustering, where k is the number of enrollments) for the embedding matching. **Configuration**: pre-trained ECAPA-TDNN; without speech enhancement; whole enrollments; Whisper, Speechbrain CRDNN, and non-speech enrollments for VAD; and sentence embedding. **Results**: With the nearest enrollment method, the resulting DER is 0.3937. With k-means, the DER was 0.3796. The difference was not stat. sig. (p =0.0596).

5.7 Different VAD Methods

We compared different VAD methods and combinations. Specifically, we used compared SpeechBrain CRDNN with vs. without non-speech enrollments. We also compared a two-stage VAD consisting of either Silero or CRDNN (with threshold of 0.9), combined with Whisper. **Configuration**: pre-trained ECAPA-TDNN; without speech enhancement; whole enrollments; sentence embedding; and nearest enrollment. **Results**: CRDNN with non-speech enrollments was stat. sig. more accurate (DER 0.3937; p =7.3020 $\times$ 10^{-9}) compared to without them (DER 0.4792). This trend persisted for a variety of different VAD thresholds. Also, when comparing Silero to CRDNN (in combination with Whisper as VAD), the former attained DER of 0.3689 and the latter of 0.3876. The difference was not stat. sig. (p =0.0989).

5.8 Discussion

All of the DERs reported are arguably high – at least 35%. This is not surprising considering the high level of background noise (from other groups in the same classroom) as well as overlapping speech (which fundamentally cannot be recognized by our diarization framework). Moreover, the fact that the teacher – for whom no audio enrollment was available in our dataset – occasionally spoke to the students resulted in another source of prediction errors.

Our experiments found benefits to using multiple VAD models, non-speech enrollments, fine-tuned ECAPA-TDNN, and subselecting enrollment audio with VAD. Further, applying Whisper facilitates simple downstream analysis to interpret who-said-what during the group collaboration. On the other hand, the configurations that we explored using speech enhancement to preprocess the audios, and voting mechanisms across the frames within each sentence, did not improve the DER.

5.9 Best-Performing Model

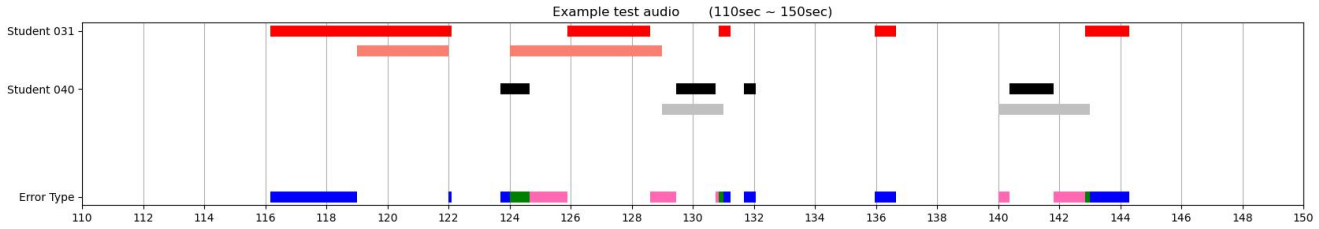

Figure 3: Speaker diarization example from our dataset. Dark colors are ground-truth, and light colors are predictions. Error Types: Pink, blue and green represent false alarm, missed detection and confusion respectively. The system achieves a DER of 0.26 on this segment.

Taking into account factors such as accuracy, time consumption and complexity, we opted for the following configuration for our framework: without enhancement; sentence embedding; fine-tuned ECAPA-TDNN; Whisper, CRDNN and non-speech enrollments as VAD; and clustering. This achieved a DER of 0.3446, consisting of False Alarm of 0.0596, Missed Detection of 0.2434 and Confusion of 0.0416.

Figure 3 shows an example of the prediction for a segment of a test audio. The figure illustrates the ground-truth speech intervals of the two speakers as well as the darization system's predictions for the interval from 110 second to 150 second of one of the test audio. In the figure, the intervals marked in red and black are the ground-truth of speaker Student 031 and Student 040 respectively. The intervals marked in light red and light gray are the corresponding predictions of each speaker. The bottom line ("Error Type") indicates whether a prediction was a false alarm, missed detection, or confusion. Correct predictions are not marked and show as blank in the Error Type row.

The model exhibits a large improvement compared to a baseline diarization method with DER of 0.7575 (configuration: without speech enhancement; frame-wise; pre-trained ECAPA-TDNN; CRDNN as VAD; and nearest enrollment). The difference is stat. sig. ($p = 4.5930 \times 10^{-15}$).

6. APPLICATION: ESTIMATING HOW MUCH EACH STUDENT SPEAKS

Given the best configuration of the Speaker Diarization Framework that we found during the previous experiments, we explored how this could be applied to real-world classroom group discussion analysis. For each individual student, the proportion of how much they speak within their group serves as an important metric to evaluate their participation in the collaboration. Additionally, for each group, the balance of speech proportion across the group members is another important indicator for teachers to assess the discussion patterns of the group as well as the role and discourse authority of each member during the discussion.

To assess the capability of the diarization system for this purpose, we calculated for each test audio in our dataset the estimated proportion of speech by each person out of the total length of group discussion in which they appeared. We then calculated the correlation (both Pearson and Spearman) between the proportions estimated by the diarizer with the proportions obtained from human annotations. We computed these correlations for Pearson and Spearman correlations of 0.5516 and 0.6208, respectively.

7. CONCLUSION

With the goal of automatically characterizing the group dynamics within classroom collaborative discussions, we have performed a systematic comparison of different design configurations of a Speaker Diarization Framework that can understand classroom speech. We assessed Diarization Error Rate on a real-world and "in-the-wild" dataset of group science discussions from middle- and high-school discussions. The best system we tried achieved a DER of around 0.34 on our test set. Moreover, the system can estimate the proportion of speech by different speakers in the group with a correlation of up to 0.62 compared to human annotations.

Future work: Instead of embedding-based diarization systems such as ECAPA-TDNN, we could use end-to-end neural models such as [31, 10, 14]. These afford the opportunity to capture simultaneous speech from multiple speakers and might achieve a better DER. With a system to identify simultaneous speech, we could then also detect automatically when one student *interrupts* another student; this could serve as a useful feedback signals to students themselves to make sure that each person's contributions are heard.

Acknowledgement

This research was supported by the NSF National AI Institute for Student-AI Teaming (iSAT) under grant DRL #2019805, and also from an NSF CAREER grant #2046505. The opinions expressed are those of the authors and do not represent views of the NSF.

8. REFERENCES

[1] Amazon. Amazon transcribe, 2021.

[2] X. Anguera, S. Bozonnet, N. Evans, C. Fredouille, G. Friedland, and O. Vinyals. Speaker diarization: A review of recent research. *IEEE Transactions on audio, speech, and language processing*, 20(2):356–370, 2012.

[3] W. Beccaro, M. A. Ramírez, W. Liaw, and H. R. Guimarães. Analysis of oral exams with speaker diarization and speech emotion recognition: A case study. *IEEE Transactions on Education*, 2023.

[4] H. Bredin. pyannote. metrics: A toolkit for reproducible evaluation, diagnostic, and error analysis of speaker diarization systems. In *INTERSPEECH*, pages 3587–3591, 2017.

[5] J. Cao, A. Ganesh, J. Cai, R. Southwell, E. M. Perkoff, M. Regan, K. Kann, J. H. Martin, M. Palmer, and S. D'Mello. A comparative analysis of automatic speech recognition errors in small group classroom discourse. In *Proceedings of the 31st ACM Conference on User Modeling, Adaptation and Personalization*, pages 250–262, 2023.

[6] S. Chen, C. Wang, Z. Chen, Y. Wu, S. Liu, Z. Chen, J. Li, N. Kanda, T. Yoshioka, X. Xiao, et al. Wavlm: Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing*, 16(6):1505–1518, 2022.

[7] H. T. Dennis Chun-Lok Fung and K. Leung. The influence of collaborative group work on students' development of critical thinking: the teacher's role in facilitating group discussions. *Pedagogies: An International Journal*, 11(2):146–166, 2016.

[8] B. Desplanques, J. Thienpondt, and K. Demuynck. Ecapa-tdnn: Emphasized channel attention, propagation and aggregation in tdnn based speaker verification. *arXiv preprint arXiv:2005.07143*, 2020.

[9] S. Dutta, D. Irvin, J. Buzhardt, and J. H. Hansen. Activity focused speech recognition of preschool children in early childhood classrooms. In *Proceedings of the 17th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2022)*, pages 92–100, 2022.

[10] Y. Fujita, N. Kanda, S. Horiguchi, Y. Xue, K. Nagamatsu, and S. Watanabe. End-to-end neural speaker diarization with self-attention. In *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 296–303. IEEE, 2019.

[11] A. Gomez, M. S. Pattichis, and S. Celedón-Pattichis. Speaker diarization and identification from single channel classroom audio recordings using virtual microphones. *IEEE Access*, 10:56256–56266, 2022.

[12] GoogleCloud. Detect different speakers in an audio recording, 2021.

[13] A. Hagen, B. L. Pellom, and R. A. Cole. Children's speech recognition with application to interactive books and tutors. *2003 IEEE Workshop on Automatic Speech Recognition and Understanding*, pages 186–191, 2003.

[14] M.-K. He, J. Du, and C.-H. Lee. End-to-end audio-visual neural speaker diarization. In *Proc. Interspeech*, pages 1461–1465, 2022.

[15] J. R. Howard. *Discussion in the college classroom: Getting your students engaged and participating in person and online*. John Wiley & Sons, 2015.

[16] S. Kelly, A. M. Olney, P. Donnelly, M. Nystrand, and S. K. D'Mello. Automatically measuring question authenticity in real-world classrooms. *Educational Researcher*, 47(7):451–464, 2018.

[17] F. Landini, J. Profant, M. Diez, and L. Burget. Bayesian hmm clustering of x-vector sequences (vbx) in speaker diarization: theory, implementation and analysis on standard tasks. *Computer Speech & Language*, 71:101254, 2022.

[18] K. Markov and S. Nakamura. Improved novelty detection for online gmm based speaker diarization. In *Ninth Annual Conference of the International Speech Communication Association*, 2008.

[19] A. M. Olney, P. J. Donnelly, B. Samei, and S. K. D'Mello. Assessing the dialogic properties of classroom discourse: Proportion models for imbalanced classes. *International Educational Data Mining Society*, 2017.

[20] T. J. Park, N. Kanda, D. Dimitriadis, K. J. Han, S. Watanabe, and S. Narayanan. A review of speaker diarization: Recent advances with deep learning. *Computer Speech & Language*, 72:101317, 2022.

[21] S. S. Pradhan, R. A. Cole, and W. H. Ward. *MyST Children's Conversational Speech*. Linguistic Data Consortium, 2021. Catalog LDC2021S05.

[22] F. Quansah. Traditional or performance assessment: What is the right way to assessing learners. *Research on Humanities and Social Sciences*, 8(1):21–24, 2018.

[23] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever. Robust speech recognition via large-scale weak supervision. In *International Conference on Machine Learning*, pages 28492–28518. PMLR, 2023.

[24] M. Ravanelli, T. Parcollet, P. Plantinga, A. Rouhe, S. Cornell, L. Lugosch, C. Subakan, N. Dawalatabad, A. Heba, J. Zhong, J.-C. Chou, S.-L. Yeh, S.-W. Fu, C.-F. Liao, E. Rastorgueva, F. Grondin, W. Aris, H. Na, Y. Gao, R. D. Mori, and Y. Bengio. SpeechBrain: A general-purpose speech toolkit, 2021. arXiv:2106.04624.

[25] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn. Speaker verification using adapted gaussian mixture models. *Digital signal processing*, 10(1-3):19–41, 2000.

[26] M. Rouvier, P.-M. Bousquet, and B. Favre. Speaker diarization through speaker embeddings. In *2015 23rd european signal processing conference (eusipco)*, pages 2082–2086. IEEE, 2015.

[27] K. Sedova, M. Sedlacek, R. Svaricek, M. Majcik, J. Navratilova, A. Drexlerova, J. Kychler, and Z. Salamounova. Do those who talk more learn more? the relationship between student classroom talk and student achievement. *Learning and Instruction*, 63:101217, 2019.

[28] Shobaki, Khaldoun, Hosom, John-Paul, and Cole, Ronald Allan. CSLU: Kids' Speech Version 1.1, Nov. 2007.

[29] R. Southwell, S. Pugh, E. M. Perkoff, C. Clevenger, J. B. Bush, R. Lieber, W. Ward, P. Foltz, and S. D'Mello. Challenges and feasibility of automatic speech recognition for modeling student collaborative discourse in classrooms. *International Educational Data Mining Society*, 2022.

[30] S. Team. Silero vad: pre-trained enterprise-grade voice activity detector (vad), number detector and language classifier. https://github.com/snakers4/silero-vad, 2021.

[31] C. Zhang, J. Shi, C. Weng, M. Yu, and D. Yu. Towards end-to-end speaker diarization with generalized neural speaker clustering. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8372–8376. IEEE, 2022.