Optimizing Speaker Diarization for the Classroom: Applications in Timing Student Speech and Distinguishing Teachers from Children

Jiani Wang
Worcester Polytechnic Institute
Worcester, United States
jwang21@wpi.edu

Shiran Dudy
Northeastern University
Boston, United States
shirdu2@gmail.com

Xinlu He
Worcester Polytechnic Institute
Worcester, United States
xhe4@wpi.edu

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

Rosy Southwell
University of Colorado Boulder
Boulder, United States
rosy.southwell@colorado.edu

Jacob Whitehill
Worcester Polytechnic Institute
Worcester, United States
jrwhitehill@wpi.edu

An important dimension of classroom group dynamics & collaboration is how much each person contributes to the discussion. With the goal of distinguishing teachers' speech from children's speech and measuring how much each student speaks, we have investigated how automatic speaker diarization can be built to handle real-world classroom group discussions. We examined key design considerations such as the level of granularity of speaker assignment, speech enhancement techniques, voice activity detection, and embedding assignment methods to find an effective configuration. The best speaker diarization system we found was based on the ECAPA-TDNN speaker embedding model and used Whisper automatic speech recognition to identify speech segments. The diarization error rate (DER) in challenging noisy spontaneous classroom data was around 34%, and the correlations of estimated vs. human annotations of how much each student spoke reached 0.62. The accuracy of distinguishing teachers' speech from children's speech was 69.17%. We evaluated the system for potential accuracy bias across people of different skin tones and genders and found that the accuracy did not show statistically significantly differences across either dimension. Thus, the presented diarization system has potential to benefit educational research and to provide teachers and students with useful feedback to better understand their classroom 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 subjects such as math, reading, and writing, but also develop broader critical thinking, communication, and collaborative learning skills (Fung et al., 2016). These skills are not limited to any particular subject but permeate the entirety of students' learning journeys. Instrumental for nurturing these skills is to incorporate discussions into the classroom, either as whole-class interactions or in small groups. Students who actively participate in these discussions tend to achieve better learning outcomes than those who do not (Howard, 2015). Moreover, a strong correlation exists between the frequency and quality of student talk during a lesson and student achievement (Sedova et al., 2019). 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" for each student is an important metric to analyze a student's learning process.

Measuring classroom speech: Given the importance of fostering effective student collaboration in group discussions, it could be beneficial for educators to automatically gauge students' behaviors in classroom group discussions, both to facilitate large-scale research studies and to provide learners with feedback. However, measuring classroom speech is challenging due to the complexity of real-world classroom dynamics. Traditional methods of assessing classroom activities, such as inviting experts into the classroom to observe and manually record discussions, or relying on survey questionnaires, cannot provide a comprehensive assessment of students' performance in group discussions (Quansah, 2018). These methods are labor-intensive, time-consuming, prone to subjective biases, and therefore may lack accuracy and objectivity. Having an automatic tool to assist the teacher in understanding each group's discussions and evaluating students' performance within each group would thus be highly valuable. Currently, deep-learning-based methods have been widely applied to various aspects of automated classroom analysis (Ramakrishnan et al., 2023; Sümer et al., 2021; He et al., 2024; Alharbi, 2023; Gazawy et al., 2023; Thomas et al., 2024). In this work, we aim to implement automated classroom speech analysis using deep learning techniques.

Speaker diarization for classroom discussion analysis: Modern deep-learning-based speech analysis algorithms offer new ways to measure both the quality and quantity of classroom speech, and they can help avoid some of the problems and risks associated with traditional assessments. In particular, speaker diarization algorithms can automatically identify "who is speaking when" (Anguera et al., 2012). Speaker diarization can assist educators in understanding students' level of participation and in assessing their communication and collaboration skills. However, to date, there is a lack of research on how speaker diarization can be deployed in noisy, real-world classrooms with the unscripted speech of children. Our paper seeks to help fill this gap. Unlike other classroom analytic methods that require specialized hardware (e.g., one LENA microphone for each child), our approach requires only a single table-top microphone for each group, making it easier to deploy and less obtrusive.

Research contribution: This paper is an extension of the paper "Speaker Diarization in the Classroom: How Much Does Each Student Speak in Group Discussions?" (Wang et al., 2024). In the previous paper, we systematically explored the design and implementation of a robust and accurate speaker diarization system capable of identifying speech segments and the corresponding speakers during classroom group discussions. Due to privacy concerns, we focused on locally deployable solutions rather than cloud-based diarization services. We presented a general speaker diarization framework and described how it could be deployed in real-world classrooms

to quantify each person's contribution to a group discussion. Building on these achievements, we extend our work in the following aspects. We employ a new dataset, the Casual Conversation Dataset, to investigate whether the embedding model performs differently across various demographics. Additionally, we incorporate another public dataset, the National Center for Research on Early Childhood Education Pre-Kindergarten Dataset (NCRECE), to validate the robustness of the core idea and demonstrate another real-world application.

The significance of this research lies in its potential to drive changes in educators' classroom management and student assessment methods. Firstly, automated speech recognition and speaker identification processes alleviate teachers' burdens in classroom supervision, 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 students' performance, promoting fairness and effectiveness in educational practices. Furthermore, since each student is a unique individual, the detailed feedback provided by the system for each student enables educators to offer personalized guidance.

Outline: In the following sections, we will first introduce the related work in this field in Section 2, and then describe the three datasets we use in this study in Section 3. Section 4 will present the proposed framework, and in Section 5 we will provide the details of the experiments and the results. Section 6 includes an evaluation of bias for the framework. Sections 7 and 8 present two real-world applications of the proposed framework.

2. Related Work

2.1. Speech Analysis of Classroom Interactions

There are numerous applications of speech processing methods in educational data mining and learning analytics. Beccaro et al. (2024) utilized speaker diarization as the core method to build a speech processing model to assess student performance and engagement during oral exams. They then examined the correlation between the emotional expressions of the students during speech and their final scores on the oral exam. Gomez et al. (2022) also employed speaker diarization for classroom analysis, which is a similar application to ours. They confronted the same challenges we faced, such us limited data with substantial noise. Instead of using deep learning methods, they addressed the problem using a physical-based model and virtual microphones. They computed the spatial information of the speakers based on speaker geometry and estimated the room impulse responses (RIRs). Ultimately, they predicted the speakers based on the cross-correlation matrix calculated from the RIRs. Olney et al. (2017) utilized SMOTEBoost to tackle the issue of imbalanced data across different categories when automatically assessing the dialogic properties of classroom discourse in real classrooms.

Cao et al. (2023) investigated the impact of automatic speech recognition (ASR) errors on the analysis of collaborative classes and provided constructive suggestions for optimizing group discourse modeling tasks. Dutta et al. (2022) proposed a translation framework that applied ASR to track the conversational speech of preschool children. Kelly et al. (2018) applied ASR to detect authentic questions in the classroom to support improvements in teaching effectiveness.

2.2. Speaker Diarization

Speaker diarization aims to automatically identify "who speaks when" within an input audio (Park et al., 2022). There are various mature methodologies to achieve this, including feature embeddings (Rouvier et al., 2015), speaker modeling (Reynolds et al., 2000; Markov and Nakamura, 2008), segmentation and clustering algorithms (Landini et al., 2022), as well as end-to-end methods (Zhang et al., 2022; Fujita et al., 2019; He et al., 2022). In recent years, an increasing number of approaches based on deep learning models have been proposed for speaker diarization. Desplanques et al. (2020) used an Emphasized Channel Attention, Propagation and Aggregation (ECAPA) deep learning model based on the Time-Delay Neural Network (TDNN) system, named ECAPA-TDNN. In this model, they applied architectural enhancements, additional skip connections, and channel attention to improve performance. Chen et al. (2022) proposed WavLM 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 during pre-training. WavLM achieves state-of-the-art performance on the CALLHOME speaker diarization benchmark. Finally, Amazon (Amazon, 2021), Google (GoogleCloud, 2021) and other companies offer cloud-based diarization services. However, for many schools, these services are unacceptable due to privacy concerns.

3. Datasets

There are three datasets involved in this study: the Sensor Immersion Dataset, which is utilized to explore the architecture and configurations of the speaker diarization framework; the Casual Conversation Dataset, which is employed to assess potential biases in the proposed framework; and the National Center for Research on Early Childhood Education Pre-Kindergarten Dataset (NCRECE), which is used in one of the real-world applications.

3.1. Sensor Immersion Dataset

As in our previous study, we used the Sensor Immersion Dataset (Southwell et al., 2022) for the diarization task here. This dataset includes both enrollment audio and test audio. 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 approximately 5 minutes long. All recordings are unscripted and contain authentic student interactions. Each audio was recorded during a group discussion involving 2 to 4 students, who were discussing how to use different sensors (temperature, moisture, CO_2, etc.) to complete a collaborative science task. Rather than providing each student with their own microphone, which was both inconvenient and arguably intrusive to both teachers and students, we used omnidirectional table-top microphones to record the audio. Due to the presence of multiple discussion groups within the same classroom simultaneously, the audio recordings contain significant noise, including non-speech noise (e.g., traffic, air conditioners) as well as non-target speech noise from students in other groups. The proportions of audio containing different numbers of simultaneous speakers are shown in Table 1.

3.1.1. Speaker Enrollment

Prior to each group discussion, the students in the group "enroll" themselves by recording a sentence of their voice and stating their name. Each enrollment audio is at least 5 seconds long

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

Table 1: Proportions of simultaneous speech from different numbers of speakers in the Sensor Immersion Dataset.

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

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

3.1.2. Annotation

All of the audios in our dataset were manually labeled for "who-spoke-when". In particular, each utterance spoken by a student was annotated with its 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.

3.1.3. Challenges

In the Sensor Immersion setting, students were divided into groups of two to four people, and each group was recorded by a table-top omnidirectional microphone. However, since all groups were in the same classroom, each microphone captured not only the voices of its own group but also the voices of the other groups. Furthermore, due to the limited amount of actual classroom data (only 32 recordings), which were reserved for testing the system's efficacy, we were unable to use this data for training or fine-tuning models.

3.2. Casual Conversation Dataset

Due to the lack of gender and skin tone labels for children in the Sensor Immersion Dataset, we used alternative datasets to assess race and gender bias. We employed the Casual Conversation Dataset published by Meta (Hazirbas et al., 2022), which is "designed to help researchers evaluate the accuracy of their computer vision and audio models across a diverse set of age, genders, apparent skin tones, and ambient lighting conditions". Gender labels in this dataset were self-reported by participants as either male or female. We used skin tone labels as a proxy for race and/or ethnicity, with the rationale that if there was bias across skin tones, there might also be bias across race/ethnicity. The skin tones in this dataset were annotated by a group of professionally trained annotators using the Fitzpatrick skin type scale (Fitzpatrick, 1988). Skin tones were classified into six levels, ranging from light to dark, categorized as Type 1 to Type 6. The videos in this dataset were recorded in the United States. Participants in the videos were asked random questions from a pre-approved list, and they provided their unscripted answers. Each video was recorded independently, featuring only one participant. The audio was clear with minimal noise. Occasionally, some audios contained the voice of the person asking the questions, but this individual did not appear in the videos.

Due to the large scale of the entire dataset, which contained a total of 3011 speakers, we selected a subset for bias evaluation. This subset comprised 78 speakers and the detailed distributions of their skin tones and genders are shown in Table 2 and Table 3. The average length of the audio was 68 seconds, with the longest being nearly 2 minutes and the shortest about 30 seconds.

Table 2: Casual Conversation Dataset Speakers Distribution by Skin Tone Type

Skin Tone Type	Speakers Number
1	8
2	17
3	17
4	2
5	12
6	22

Table 3: Casual Conversation Dataset Speakers Distribution by Gender

Gender	Speakers Number
Female	44
Male	34

3.3. National Center for Research on Early Childhood Education Pre-Kindergarten Dataset

To explore the potential application of the proposed framework to a different age group, we investigated its performance in distinguishing teachers' speech from children's speech using the National Center for Research on Early Childhood Education Pre-Kindergarten Dataset (NCRECE) (Pianta and Burchinal, 2016; Pianta et al., 2017). This dataset is widely used in preschool classroom analysis (Whitehill and LoCasale-Crouch, 2023; Yang et al., 2023).

Unlike the Sensor Immersion Dataset, NCRECE primarily captures interactions in early childhood classrooms, focusing on a younger age group with children aged around 4 years old. Consequently, teachers are the primary speakers in these classrooms, accounting for over 75% of the speech. Moreover, students in this age group often have low awareness of rules and exhibit more problematic behaviors, such as suddenly shouting or engaging in other disruptive actions (Luczynski and Hanley, 2013). As a result, the audios are noisy, containing significant background noise and instances of overlapping speech from multiple people.

3.3.1. Annotation

We randomly selected a subset of the NCRECE dataset and annotated based on text and audio respectively.

Text-based Annotation: We first generated transcripts using Whisper large-v2, and then split them into individual sentences. Annotators determined whether each sentence originated from a teacher or a student based solely on the textual information of the sentence, without listening to the audio or considering the surrounding context.

Audio-based Annotation: Following the initial step in Text-based Annotation, we utilized Whisper to generate transcripts and obtained the corresponding timestamps for each sentence. Then we extracted the audio clips of these sentences from the original audios. Finally, annotators listened to each clip and labeled the speaker's role (teacher or student) without any access to textual information. If multiple speakers were talking simultaneously, the annotators assigned the label based on the dominant speaker.

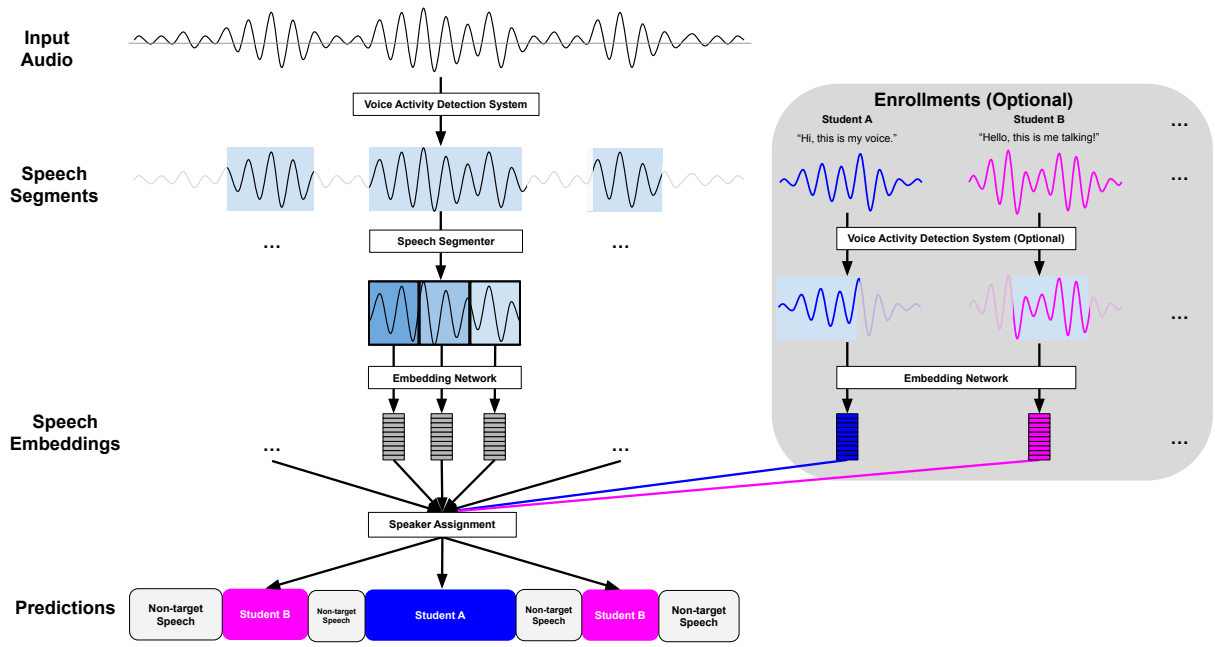

Figure 2: Speaker diarization framework

4. Speaker Diarization Framework

In our previous work, we outlined the general framework used to perform speaker diarization, including various design variants we explored (Wang et al., 2024). The inputs to our speaker diarization system are consistently (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 through several phases (see Figure 2), as described in the following subsections.

4.1. Speech Enhancement

Unconstrained audio recordings from table-top microphones in school classrooms, where multiple simultaneous discussions among different student groups occur, can be highly noisy. Hence, as an optional initial step, non-speech noise can be filtered out using a speech enhancement system, for example, the SpeechBrain WaveformEnhancement, to improve speech quality (Ravanelli et al., 2021).

For enrollments, we applied the enhancer directly to the original enrollment audios. For test audio, we applied it only to the active speech segments obtained from the Voice Activity Detection (VAD) system as described in the next section.

4.2. Voice Activity Detection System

Many automatic speech applications use Voice Activity Detection (VAD) to identify segments of input audio that contain speech. In our diarization system, segments identified as not containing speech by the VAD system were excluded from further processing and immediately classified as "no speech" in our diarization system. Inevitably, some errors occur where segments with speech

were incorrectly identified as "no speech". These errors were reflected in the DER calculation as missed detections.

We applied the VAD system differently to enrollments and test audio, as described below.

4.2.1. Enrollments

The enrollment audios in our study were often very noisy and included segments without speech. To enhance the quality of the enrollment audio, we explored applying a VAD system to select the most useful parts. Specifically, we used the SpeechBrain Convolutional Recurrent Deep Neural Network (CRDNN) VAD model (Ravanelli et al., 2024) to identify speech segments with the highest probability of containing speech.

4.2.2. Test audio

For processing the test audio, we explored three different methods to detect non-target speech: Whisper ASR, non-target speech enrollments, and a secondary VAD system (either SpeechBrain CRDNN or Silero (Silero Team, 2024)). We also tested combinations of these approaches, as described below.

Whisper: In pilot experiments, we found that Whisper (Radford et al., 2023), an automatic speech recognition system, could effectively serve as a VAD method. Whisper generated a list of start and end timestamps for spoken sentences in any input audio, along with the estimated transcript. These timestamps generally corresponded well to periods of speech throughout the entire test audio. For our application, we ignored the transcript and utilized only the timestamps. Using Whisper in this way was often beneficial for various downstream applications (e.g., inferring who said what in a discussion group). Throughout this paper, Whisper large-v2 is used unless a different version is specified.

Non-target speech enrollments: As an additional way to detect non-target speech moments in the test audio, we compared extracted speech embeddings to an embedded audio of background noise. For each audio, we extracted a segment ($\geq$ 5 seconds) containing no speech and only background noise. This segment served as the "non-target speech enrollment". Subsequently, this enrollment was treated on par with the enrollments of candidate speakers. Hence, when using this approach, the enrollment set for each test audio comprised enrollments of all speakers present as well as the non-target speech enrollment.

Secondary VAD method: We experimented with combining Whisper ASR as a first-stage VAD method with a secondary VAD model (either SpeechBrain CRDNN or Silero). For each test audio, we first applied Whisper for initial VAD, yielding intervals containing at least one speaker. Subsequently, based on the start and end times of these intervals, we extracted corresponding segments from the test audio, referred to as Whisper-segments. These Whisper-segments then served as inputs for the secondary VAD model (either SpeechBrain CRDNN or Silero). In the second stage, the secondary VAD model processed the Whisper-segments and computed the probability of speech for each segment, denoted as $pred_{speech}$. If $pred_{speech}$ exceeded a predefined threshold, the segment was identified as containing at least one speaker, and its corresponding intervals were saved for further embedding calculations. Conversely, if $pred_{speech}$ fell below the threshold, the segment was immediately classified as "no speech".

VAD threshold: For most experiments, we set the SpeechBrain CRDNN threshold to 0.8798, selected through the following process. We calculated the averaged speech probability for segments contain at least one speaker, referred to as $prob_{act}$, and the average probability for seg-

ments without any speaker, referred to as $prob_{de}$. The threshold for CRDNN was then defined as the average of $prob_{act}$ and $prob_{de}$.

When utilizing CRDNN for VAD in the test audio, CRDNN output a frame-level probability for each audio chunk. We employed two approaches for using the VAD threshold. On the one hand, we calculated the average probability across all frames in a segment and compared this value to the threshold to determine if the segment contained speech. On the other hand, instead of applying only one threshold, we applied two to detect the start and end times of speech, named *active threshold* and *inactive threshold*. When the probability at a certain timestamp exceeded the active threshold, speech was considered to have started. Subsequent probabilities only needed to exceed the inactive threshold to maintain the detection of speech. If at any point, the probability dropped below the inactive threshold, speech was considered to have ended.

Combination of VAD model and non-target speech enrollments: When employing both VAD model(s) and non-target speech enrollments, we first utilized VAD model(s) to identify intervals containing at least one speaker. Subsequently, these resulting intervals were then used to extract the corresponding audio segments and obtain embeddings. For each embedding, we calculated the cosine similarity not only with embeddings of all candidate speakers' enrollments but also with the embedding of a non-target speech enrollment. If the cosine similarity with the non-target speech enrollment embedding was the highest, the segment was labeled as "no speech".

4.3. Speech Segmenter

For the detected speech sentences (represented in blue in Figure 2), we might either process each sentence as a whole or divide them into fixed-length frames. When splitting into frames, we used a frame width of 2 seconds and a step size 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, where embeddings from the same speaker are close together and embeddings from different speakers are far apart. For our model architecture, we used the Emphasized Channel Attention, Propagation and Aggregation in Time Delay Neural Network (ECAPA-TDNN) model (Desplanques et al., 2020), which, as of 2024, is state-of-the-art. We used either an off-the-shelf (pre-trained) or a fine-tuned version of ECAPA-TDNN (see below) to extract embeddings. ECAPA-TDNN can accept variable-length audio segments as input, allowing it to be applied to either an entire Whisper-segment or individual frames extracted from within a Whisper-segment. In addition to processing each segment of the test audio, we also computed embeddings for each enrollment audio clip. Although we initially considered using WavLM, pilot experiments showed its performance was inferior, so we did not pursue this approach further.

For fine-tuning ECAPA-TDNN, we used a variety of public datasets containing children's speech, specifically the CUKids (Hagen et al., 2003) (118 hours from 1354 speakers), CSLUKids (Shobaki, Khaldoun et al., 2007) (98 hours from 1118 speakers), and MyST (Pradhan et al., 2023) (435 hours from 1300 speakers) datasets. Jointly, these datasets covered students aged 5-16 and included 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. Speaker Assignment

Given the embeddings extracted from each processed segment of input audio, the next step was to assign speakers to these embeddings. Enrollments were optional in this process. When available, they allowed us to match embeddings to specific speakers and identify who they were; otherwise, speakers were labeled generically, such as $Speaker_i$. In the group discussions in our dataset, multiple students occasionally spoke simultaneously. However, since these cases occurred rarely ($< 2\%$ of the time), we ignored such possibilities and always assigned a speech segment to a single speaker. There were two design questions for this process: when enrollments were available, whether to assign embeddings to speakers using a "nearest enrollment" vs. clustering; and the level of granularity at which the assignment was made.

4.5.1. Nearest Enrollment vs. Clustering

A common approach in speaker diarization is to calculate the cosine similarity between the embedding of each test segment and each enrollment embedding, assigning 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. In contrast to the nearest enrollment methods, clustering algorithms can sometimes harness the entire trajectory (over a classroom audio) of embeddings to find a better centroid to represent each speaker (rather than just using the enrollment as the centroid). Moreover, some algorithms, such as Density-Based Spatial Clustering of Applications with Noise (DBSCAN), can partition the embedding space into speaker groups more flexibly than nearest-neighbors algorithms. In addition, speaker enrollments may not always be available. In such cases, clustering the embeddings, rather than assigning them based on nearest enrollment, is essential. Therefore, we explored various clustering methods, including k-means, agglomerative clustering, mean shift clustering and DBSCAN.

Clustering methods fall into two categories: those that require specifying the target number of clusters (in our case, k-means and agglomerative clustering) and those that can estimate the number of clusters automatically (in our case, mean shift clustering and DBSCAN). Given that we know the number of students in each discussion group, we can set the number of clusters equal to the number of speakers. In contrast, for the latter two clustering methods, we do not initialize the target number of clusters prior to the experiment. After clustering, the Hungarian algorithm is used to find the optimal matching between clusters and speakers. 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 the audio is originally split into frames, we can either assign each frame to a speaker, or (if using Whisper as VAD model) aggregate the frames within each sentence to assign a single speaker to each sentence through one of several possible voting mechanisms. Alternatively, we can compute an embedding for each entire sentence as segmented by Whisper and directly assign each sentence to a speaker. Embedding each frame has the possible advantage of capturing "purer" speech segments, as the likelihood of including speech from multiple speakers is reduced. On the other hand, analyzing each sentence as a whole leverages longer speech segments and the linguistic structure recognized by Whisper (since it performs speech recognition).

To aggregate the frames within each sentence, we have the following three voting mechanisms.

Majority Vote: Given a sentence of speech comprising frames $f_1, \ldots, f_n$, the embedding model computes embeddings $e_1, \ldots, e_n$ for each frame. Then, the cosine similarity between each embedding i and the enrollment embedding of each candidate speaker is calculated; the speaker with the highest cosine similarity s_i is selected as the predicted speaker p_i for that frame. Across all n frames within a sentence, we tally the occurrences of each speaker in the predictions, and the speaker with the highest frequency is chosen as the final prediction for the sentence.

When applying the Majority Vote method, we observed instances where certain frames had the same prediction, but their corresponding cosine similarities with the candidate speakers varied significantly. To better leverage these cosine similarity values, we developed 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 the cosine similarities for each candidate speaker across all frames. The speaker with the maximum summed similarity is chosen as the prediction for the entire sentence. Here, the weight of each frame is the product of its cosine similarity and frame length, allowing individual frames to contribute variably to the final prediction.

Argmax Vote: This method focuses on the maximum cosine similarity value. 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. Speaker Diarization Framework: Experiment & Results

We conducted experiments on different design configurations of the speaker diarization framework presented in Section 4 to determine which works the best. The experiments in this section used the Sensor Immersion Dataset.

5.1. Evaluation Metric

In this work, we employ the diarization error rate (DER) and the Spearman's Rank Correlation Coefficient (SCC) to guide the architecture and configurations search of the framework.

5.1.1. Diarization Error Rate

We evaluate accuracy using diarization error rate (DER) (Bredin, 2017), which measures the fraction of the total audio length in which the *set* of speakers (as multiple people may speak simultaneously) was incorrectly inferred by the model. This is computed as:

$$\text{DER} = \frac{\text{false alarm} + \text{missed detection} + \text{speaker confusion}}{\text{total length of the audio}}$$

where false alarm represents the length of the duration where no one was speaking but the model believed someone was, missed detection is the opposite, and speaker 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 speaker confusion. For example, if speakers A and B were talking, but the model identified only speaker A, this segment would be considered

a speaker confusion. Since our speaker diarization framework always assigns speech segments to individual speakers, it will always be penalized in DER for any segment where ground-truth involves multiple speakers.

To compute DER over our entire dataset, we calculate DER for each test audio individually, then take the weighted average of the DERs across all test audio with weights based on the length of each test audio. For significance testing between different diarization methods, we use paired t-tests across the 32 test audio.

The DER of the baseline model is 0.7575. The configurations of the baseline model include: without speech enhancement, frame-based segmentation, pre-trained ECAPA-TDNN, CRDNN for VAD, and nearest enrollment for speaker assignment.

5.1.2. Correlation Coefficients

We employ Pearson Correlation Coefficient (PCC) and Spearman's Rank Correlation Coefficient (SCC) to measure the relationship between each speaker's predicted speech proportion and human annotations for estimating how much each student speaks (Section 7).

5.2. Speech Enhancement

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

5.3. Subselecting Enrollment Audio with VAD model

We compared using the whole enrollment audio to compute the enrollment embedding for each speaker with 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" part of the enrollment audio, since pauses and volume variations may occur when recording, meaning the original enrollment may not contain speech in every frame. Therefore, we aimed to select the proportion of the original enrollment that had the highest probability of containing speech to better reflect the speaker's voice features. In particular, we selected the "best" fixed-length segment (2, 4, 8, 16, or 32 seconds) within each enrollment audio, based on the speech probability output by the SpeechBrain CRDNN. First, we used the CRDNN to obtain the frame-level posterior probabilities for the original enrollment audio, where a higher probability indicated a higher likelihood of speech presence. We selected the maximum probability and recorded its corresponding timestamp as T_{peak}. Centered on T_{peak}, we extracted a segment of specified length as the "best" part of the enrollment. Finally, we extracted an enrollment embedding based on only this portion of the enrollment audio. **Configurations**: without speech enhancement; Whisper, SpeechBrain CRDNN, and non-target speech enrollments for VAD; sentence embedding; pre-trained ECAPA-TDNN; and nearest enrollment. **Results**: Using 4-second segmented enrollments achieved a DER of 0.3745 which is statistically significantly better ($p = 0.0073$) than the model using whole enrollments, which had a DER of 0.3937.

5.4. Different VAD Methods

We compared different VAD methods and combinations. Specifically, we compared SpeechBrain CRDNN with vs. without non-target speech enrollments. We also evaluated a two-stage VAD system that used either Silero or CRDNN (with threshold of 0.9), in combination with Whisper. **Configurations**: without speech enhancement; whole enrollments; sentence embedding; pre-trained ECAPA-TDNN; and nearest enrollment. **Results**: CRDNN with non-target speech enrollments was statistically significantly more accurate (DER 0.3937; $p = 7.3020 \times 10^{-9}$) compared to without them (DER 0.4792). This trend held across various VAD thresholds. Additionally, when comparing Silero and CRDNN (each combined with Whisper as the VAD model), Silero achieved a DER of 0.3689, while CRDNN achieved a DER of 0.3876. However, this difference was not statistically significant ($p = 0.0989$).

5.5. Frame-based vs. Sentence-based Prediction

We compared frame-based assignment with four sentence-based assignment methods: sentence embedding, Majority Vote, Argmax Vote, and Weighted Vote. **Configurations**: without speech enhancement; whole enrollments; Whisper, SpeechBrain CRDNN, and non-target speech enrollments for VAD; pre-trained ECAPA-TDNN; and nearest enrollment.

Hyperparameter selection: The frame-based approach requires selecting several hyperparameters. We employed 5-fold cross validation to select these hyperparameters, which consisted of *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: we first divided the 32 test audio into 5 groups to conduct 5 sub-experiments. For each sub-experiment, we applied the diarization pipeline to 4 groups of test audio 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 audio. After completing the 5 sub-experiments, we obtained 5 pairs of best hyperparameters, all of which were identical: *windowSize*=2 and *stepSize*=0.75. Consequently, we selected this pair of hyperparameters for all frame-based experiments.

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

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

5.6. 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 with its fine-tuned version (see Section 4.4) in terms of DER. **Configurations**: without speech enhancement; whole enrollments; Whisper, SpeechBrain CRDNN, and non-target speech enrollments for VAD; sentence embedding; and nearest enrollment. **Results**: The fine-tuned model achieved a DER of 0.3577, which is statistically significantly better ($p = 0.0007$) than the pre-trained one, which had a DER of 0.3937.

5.7. Nearest Enrollment vs. Clustering

In this experiment, we compared the effectiveness of the nearest enrollment method and clustering methods for assigning speakers. **Configurations**: without speech enhancement; whole enrollments; Whisper, SpeechBrain CRDNN, and non-target speech enrollments for VAD; sentence embedding; and fine-tuned ECAPA-TDNN. **Results**: With the nearest enrollment method, the resulting DER is 0.3577. With k-means, the DER was 0.3446. With agglomerative clustering, the DER was 0.3746. The difference between agglomerative clustering and k-means, as well as the difference between k-means and the nearest enrollment method, were both statistically significant (p =0.0049 and p =0.0138 respectively). However, the difference between agglomerative clustering and the nearest enrollment method was not statistically significant (p =0.0538).

For DBSCAN and mean shift clustering, although initializing the number of clusters is not mandatory in these methods, other hyperparameters still need to be optimized. We attempted to identify a hyperparameter that would be universally applicable across all classroom audios with varying numbers of speakers, but in practice, we could not find such a hyperparameter. The resulting number of clusters significantly deviated from the actual values, with either only one cluster being formed or each embedding being isolated into its own cluster. Therefore, we conclude that clustering methods that automatically estimate the number of clusters are not suitable for the dataset used in this work.

5.8. Running Time Test

We tested the running time of the proposed framework, as well as the running time of its main components. Additionally, since Whisper is the most time-consuming module within the entire framework, we compared different versions of the Whisper model in terms of both running time and DER. The hardware configurations we used were as follows: GPU: NVIDIA A100-PCIE-40GB, Memory: 39.38 GB, CPU: x86_64, CPU Clock Speed: 3100.00 MHz. The results are shown in Table 4.

Table 4: Results for Time Test

Whisper Version	**Time (hour:minute:second)**						**DER**
	Whisper	CRDNN	Embedding Model	Speaker Assignment	DER Calculation	Total	
Medium	0:13:59	0:04:52	0:02:01	0:00:17	0:00:25	0:21:34	0.3519
Large-v2	0:18:29	0:05:25	0:02:12	0:00:30	0:00:54	0:27:30	0.3448
Large-v3	0:18:13	0:04:56	0:01:49	0:00:16	0:00:26	0:25:40	0.3441

5.9. Discussion

All of the DERs reported are arguably high which 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 enrollment audio was available in our dataset – occasionally spoke to the students resulted in another source of prediction errors.

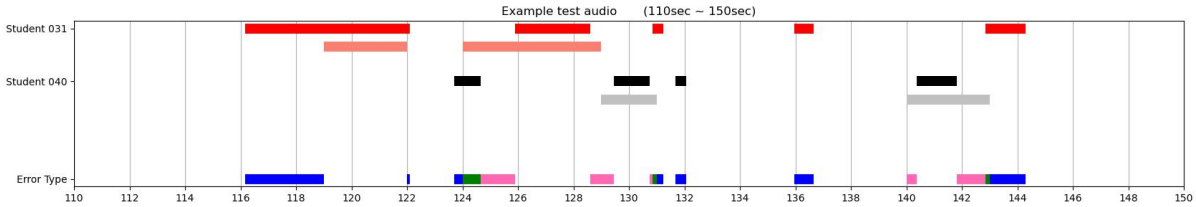

Figure 3: Speaker diarization example from our dataset. Dark colors are ground-truth, and light colors are predictions. Error Types: Pink, blue and green represent false alarm, missed detection and speaker confusion respectively. The system achieves a DER of 0.26 on this segment.

Our experiments found benefits in using multiple VAD models, non-target speech enrollments (Section 5.4), fine-tuned ECAPA-TDNN (Section 5.6), and subselecting enrollment audio with a VAD model (Section 5.3). Furthermore, applying Whisper (Section 5.4) facilitates simple downstream analysis to interpret who-said-what during group collaboration. On the other hand, the configurations we explored that used speech enhancement (Section 5.2) to preprocess the audio, and voting mechanisms across the frames within each sentence (Section 5.5), did not improve the DER.

To understand the differences in prediction results among various methods, particularly focusing on commonalities and unique errors, we selected the audio that exhibited the maximum difference in DER between the baseline model and the sentence-based best-performing model for comparison. The DER of this audio was 0.9638 when applied the baseline model, while it decreased to 0.1670 when applying the sentence-based best-performing model. By comparing these two results, we conclude that:

1. Our best model is much more conservative in what it deems to be "speech".

2. Our model still makes lots of mistakes, but it does have some aggregate sense of "Who is talking when".

5.10. Best-Performing Models

Taking into account factors such as accuracy, time consumption and complexity, we selected two best-performing models for our framework: one frame-based and one sentence-based. The other configurations of the two models are the same: without speech enhancement; whole enrollments; Whisper, CRDNN and non-target speech enrollments for VAD; fine-tuned ECAPA-TDNN; and clustering.

The frame-based best performing model achieved a DER of 0.3412, consisting of false alarm of 0.0698, missed detection of 0.2246 and speaker confusion of 0.0468; while the sentence-based best performing model achieved a DER of 0.3446, consisting of false alarm of 0.0596, missed detection of 0.2434 and speaker confusion of 0.0416. The difference between these two best models was not statistically significant (p =0.3711).

Figure 3 shows an example of the prediction for a segment of a test audio (applied to the sentence-based best-performing model). The figure illustrates the ground-truth speech intervals of the two speakers as well as the diarization system's predictions for the interval from second 110 to second 150 of one of the test audios. In the figure, the intervals marked in red and

black represent the ground-truth for speakers Student 031 and Student 040 respectively. The intervals marked in light red and light gray are the corresponding predictions for each speaker. The bottom line ("Error Type") indicates whether a prediction is a false alarm, missed detection, or speaker confusion. Correct predictions are not marked and appear blank in the "Error Type" row.

Both models exhibit a significant improvement compared to a baseline diarization method, which had a DER of 0.7575 (configurations: without speech enhancement; frame-based; pre-trained ECAPA-TDNN; CRDNN for VAD; and nearest enrollment). The improvements are both statistically significant (frame-based: $p = 1.1385 \times 10^{-15}$; sentence-based: $p = 4.5930 \times 10^{-15}$). The experimental results of all the other configurations are available in Figure 4 at the end of this paper.

6. Bias Evaluation

When using a speaker diarization system to analyze classroom speech, it is vital that the system can recognize speech from individuals of different demographics (e.g., gender, race & ethnicity) with equal accuracy. Therefore, we estimate how much bias there is in our diarization system. Since the core module of our diarization framework is the speaker embedding model — we use the ECAPA-TDNN (Desplanques et al., 2020) in particular — we focus our analysis on this component.

To verify whether the embedding model performs differently across various groups, we investigate its performance on individuals of different races and genders. The dataset used in this study is the Casual Conversation Dataset, which is introduced in Section 3.2.

6.1. Skin Tone Bias Evaluation

For each skin tone, we examined whether the distance between embeddings of the same speaker was smaller than the distance between embeddings of different speakers with the same skin tone.

We used the following procedure: Assume there are n speakers with a skin tone of Type X. We first obtained the set of *active segments* for each speaker i using Whisper and CRDNN as the VAD components to detect the intervals containing speech. From these active segments, we extracted speaker embeddings using ECAPA-TDNN to yield the *embedding set* for that speaker. For each speaker i, we randomly selected 3 embeddings whose intervals were at least 1 second long from their embedding set. We designated 1 of these 3 embeddings as the main embedding, denoted E_{im}, and the other two as comparison embeddings denoted E_{i1} and E_{i2}. We then selected another speaker with the same skin tone, for example, j, and chose 3 embeddings E_{j1}, E_{j2} and E_{j3} from their embedding set. Next, we calculated the distances between the main embedding E_m and the other five embeddings to obtain D_{mi1}, D_{mi2}, D_{mj1}, D_{mj2} and D_{mj3}. We then computed the average of D_{mi1} and D_{mi2} to represent the distance between embeddings of the same speaker, denoted D_{same}. Similarly, we calculated the average of D_{mj1}, D_{mj2} and D_{mj3} to represent the distance between embeddings of different speakers, denoted D_{diff}. If $D_{same} < D_{diff}$, we considered that in this pair of speakers, the similarity between embeddings of the same speaker was higher. Finally, we computed the probability, over all speaker pairs $i \neq j$, that $D_{same} < D_{diff}$.

To reduce variance, we repeated this procedure 25 times and computed the average probability. The results are shown in Table 5 (a). These proportions did not exhibit an increasing

or decreasing trend based on skin tone, and the proportions were not statistically significantly different from each other ($\chi^2(5) = 0.7363, \quad p = 0.9809$).

Additionally, we selected segments with a minimum length of 5 seconds to conduct the same experiments as above, and the results are presented in Table 5 (b). Similarly, the proportions did not display an increasing or decreasing trend with skin tone, and the proportions were not statistically significantly different from each other ($\chi^2(5) = 0.4435, \quad p = 0.9940$).

Table 5: Bias Evaluation for Skin Tone Types

(a) Segment Length = 1 second

Skin Tone Type	Average Correct Pairs	Total Pairs	Accuracy
1	53.12	56	94.86%
2	244.60	272	89.93%
3	260.52	272	95.78%
4	1.52	2	76.00%
5	125.96	132	95.42%
6	424.12	462	91.80%

(b) Segment Length = 5 seconds

Skin Tone Type	Average Correct Pairs	Total Pairs	Accuracy
1	53.44	56	95.43%
2	258.84	272	95.16%
3	261.76	272	96.24%
4	1.60	2	80.00%
5	128.72	132	97.52%
6	427.92	462	92.62%

6.2. Gender Bias Evaluation

In this experiment, we utilized the same process to obtain the embedding set for each speaker. For each gender, we employed the same strategy to select embeddings, calculate D_{same} and D_{diff}, and compute the number of pairs that met the criterion of $D_{same} < D_{diff}$. We repeated the experiment 25 times, calculating the average number of pairs and proportions. We also conducted two sets of experiments by selecting intervals with at least 1 second in length and intervals with at least 5 seconds in length. The results are presented in Table 6 (a) and (b) respectively, and the proportions for male and female speakers were not statistically significantly different from each other($\chi^2(1) = 0.2586, \quad p = 0.6111$ for segment = 1 second; $\chi^2(1) = 0.0350, \quad p = 0.8516$ for segment = 5 seconds).

Table 6: Bias Evaluation for Genders

(a) Segment Length = 1 second

Gender	Average Correct Pairs	Total Pairs	Accuracy
Female	1774.00	1892	93.76%
Male	1031.28	1122	91.91%

(b) Segment Length = 5 seconds

Gender	Average Correct Pairs	Total Pairs	Accuracy
Female	1788.36	1892	94.52%
Male	1068.24	1122	95.21%

7. Application: Estimating How Much Each Student Speaks

Given the two best configurations of the speaker diarization framework identified in the previous experiments, we explored how they 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 for evaluating their participation in the collaboration. Additionally, for each group, the balance of speech proportions between group members is another important indicator that teachers can use 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 the Sensor Immersion Dataset, the estimated proportion of speech for each person relative to the total length of the group discussion in which they appeared. We then calculated the correlation (both Pearson and Spearman) between the proportions estimated by the proposed diarization system and the proportions obtained from human annotations. The results are shown in Table 7.

Table 7: Results of estimating how much each student speaks

Model Name	DER	PCC	SCC
Baseline Model	0.7575	0.2363	0.3091
Frame-based Best-performing Model	0.3412	0.4497	0.5171
Sentence-based Best-performing Model	0.3446	0.5516	0.6208

Based on the experimental results, we observed that the frame-based best-performing model did not perform as well in terms of SCC in this application compared to the sentence-based best-performing model. Therefore, to explore the reason for this phenomenon, we investigated the correlation between DER and SCC. We calculated the SCC between the DER results for the diarization task and the corresponding SCC results for the application. The result (-0.5756) indicates that there is only a weak relationship between the DER outcomes of the diarization

task and the SCC results in this application. Hence, some methods may achieve better DER but worse SCC.

Additionally, the sentence-based model holds greater potential value in practical applications. Since it generates one prediction for each sentence, it better preserves the integrity of the sentences, making it advantageous for subsequent applications, particularly for content-related analyses, for example, investigating the proportion of questions asked by students.

8. Application: Identifying Teachers' and Children's Speech

Many studies on classroom teaching patterns rely on obtaining audio recordings of teachers in the classroom as a research basis. For example, Mahmood (2021) investigated what tone or intonation teachers should use to enhance communication efficiency with students in online classrooms, and Bao (2020) analyzed what speaking style teachers should use to help students better capture essential lecture points in writing. These studies typically require prior knowledge of the segments in which teachers speak, and then they analyze the speaking style of these segments. Therefore, solving the problem of identifying which segments in classroom audio correspond to teachers' speech becomes fundamental to analyzing classroom teaching patterns. In this section, we investigate whether our diarization framework can segment teachers' and children's speech using a dataset from the NCRECE (see Section 3.3). In particular, we will first show how to use the proposed framework to solve the problem of distinguishing between teachers' and children's speech. Then, we will analyze the core of the speaker diarization framework – the speaker embedding model – in terms of how well it can distinguish between teachers' and children's speech segments. At the same time, our analysis will highlight how our framework and speaker embedding model generalize to scenarios with young children (around 4 years old).

8.1. Clustering-based Approach

The NCRECE dataset does not contain any enrollments of children's or teachers' speech. We thus apply our speaker diarization framework without these enrollments. Therefore, instead of matching the speaker embedding of each segment to an enrolled person, we must use clustering to group different speech segments into identities based on the similarity of these embeddings. The exact number of children who are present at any given moment in the videos in the NCRECE dataset is not annotated, and in general, it is difficult to determine since sometimes a child may wander out of the camera's view or microphone's detection range. Due to age-related differences in speech, the distinction between teachers' and children's voice is generally greater than the distinctions between voices of students within the same age group (Harrington et al., 2007). Therefore, we can treat all students as a single entity, meaning that the audio data will effectively contain only two speakers: one being the teacher, and the other representing all students as a collective, i.e., non-teacher. We thus set the target number of clusters as 2, where the larger cluster (with more assigned speech segments) is assumed to belong to the teacher and the smaller cluster is assumed to belong to the children.

We applied k-means as the clustering method and used the audio-based labels as the ground truth. We compared the performance of the pre-trained ECAPA-TDNN model and the fine-tuned ECAPA-TDNN model. The configurations of the remaining components were: Whisper as the VAD model, sentence-based, without speech enhancement.

Additionally, we calculated the proportion of speech by teachers and children, and used SCC

to evaluate the relationship between the predicted proportions and human annotations, following the method described in Section 7. The SCC results showed a medium relationship between the predicted proportion and the human annotations, indicating that the model can estimate the speaking proportions for each individual. The results are shown in Table 8.

Table 8: Clustering Results for All Sentences

Embedding Model	DER	SCC
Pre-trained	0.3922	0.6911
Fine-tuned	0.3764	0.7155

8.2. Analysis: Accuracy of distinguishing teacher from student segments using speech embeddings

Since the core of the speaker diarization framework is the speaker embedding extractor, we analyzed the accuracy of ECAPA-TDNN in distinguishing between speech segments of teachers and children.

In our analysis, we first selected embedding groups, each consisting of two embeddings from the teacher and one embedding from the children. For each group, we calculated the cosine similarity between embeddings from the same speaker as S_{same} and that between different speakers as S_{diff}. We then calculated the proportion of groups for which $S_{same} > S_{diff}$. This proportion served as our accuracy metric for the speaker embedding system.

Using this procedure, we explored two questions: (1) Does the ECAPA-TDNN embedding model that was fine-tuned on child speakers give higher accuracy than the pre-trained ECAPA-TDNN? (2) Could obtaining multiple enrollments from the teacher improve the accuracy of distinguishing the teacher's speech from that of children? The motivation here is that, due to changes in background noise, microphone placement, etc., a teacher's voice might change over time, and by comparing to a teacher enrollment collected more recently, a diarization system could potentially achieve higher accuracy.

8.2.1. Nearest Teacher's Speech VS. Randomly Select Teacher's Speech

We have two strategies for selecting a teacher's speech for comparison. The first method involves randomly selecting two sentences, provided that the speaker is the teacher. These sentences could originate from any time during the session. The second method selects the two sentences temporally closest to the current student sentence. Since non-speech noise may fluctuate during the class, for instance, when the teacher plays a video in class or when an emergency vehicle passes by with a siren outside, we believe that selecting adjacent sentences helps maintain consistent background noise, aiding the system in distinguishing between the speech of teachers and children.

The other configurations for this experiment were: applying Whisper large-v2, using text-based labels. The results are shown in Table 9.

8.2.2. Fine-tuned ECAPA-TDNN VS. Pre-trained ECAPA-TDNN

Similar to the configurations of the proposed framework, we explored the performance of different embedding models in this task, including pre-trained ECAPA-TDNN and fine-tuned ECAPA-

Table 9: Accuracy Results for Different Teacher's Speech

Teacher's Speech	Embedding Model	Accuracy
Randomly	Pre-trained	0.5965
	Fine-tuned	0.6015
Nearest	Pre-trained	0.6917
	Fine-tuned	0.6736

TDNN. The results indicate that the fine-tuned ECAPA-TDNN does not demonstrate a significant advantage in this task and the performance of the pre-trained ECAPA-TDNN even surpasses that of the fine-tuned ECAPA-TDNN. The results are shown in Table 10.

Table 10: Accuracy Results for Different Embedding Models

Embedding Model	Accuracy	
	Text-based Label	Audio-based Label
Pre-trained	0.6917	0.6702
Fine-tuned	0.6736	0.6300

We attribute the phenomenon where the fine-tuned ECAPA-TDNN did not outperform the pre-trained ECAPA-TDNN, unlike in the previous group discussion task (Section 5.6), to differences in the scenarios of the two tasks. Although both datasets were collected in classrooms, the Sensor Immersion Dataset primarily focused on group discussions, with students as the main speakers and most participants being middle- to high-school students. In contrast, this task focuses on younger students (around 4 years old), with the teacher as the primary speaker. Additionally, the data used to fine-tune the model mainly consisted of students aged 5 to 16, which differs from this task.

8.3. Discussion

In this experiment, we were also curious whether longer sentences necessarily yield higher accuracy. Thus, for each sub-experiment with different configurations, we also calculated the SCC between the accuracy and the number of words in the sentence to investigate the relationship between them. The results indicate that, despite the large variance in SCC obtained from different experiments, all SCC values are positive, indicating a moderate correlation between accuracy and sentence length.

During the experiment, we found that the timestamps generated by Whisper large-v2 were not always precise, especially for long audio segments. Therefore, we tried WhisperX, which was proposed by Bain et al. (2023) aimed to address this problem. The other configurations were: audio-based label, nearest teacher's speech. The results are shown in Table 11. We found that WhisperX did not outperform Whisper large-v2. One possible explanation is the different segment lengths recognized by the two models. WhisperX partially corrected for the timestamps of certain sentences. However, it also tended to split some originally continuous sentences into individual words, each with its own timestamp. This resulted in WhisperX producing a

higher number of very short segments. Statistical analysis revealed that in experiments where we applied the pre-trained ECAPA-TDNN and the audio-based label, segments lasting between 0 and 1 second accounted for approximately 20.22% of the recognized segments in Whisper large-v2, compared to 29.33% in WhisperX.

Table 11: Accuracy Results for Different Whisper Models

Whisper Version	Embedding Model	Accuracy
Large-v2	Pre-trained	0.6702
	Fine-tuned	0.6300
X	Pre-trained	0.6568
	Fine-tuned	0.6307

9. Conclusion

With the goal of automatically characterizing group dynamics within classroom collaborative discussions, we have conducted a systematic comparison of different design configurations of a speaker diarization framework capable of understanding classroom speech. We assessed DER on a real-world and "in-the-wild" dataset of group science discussions from middle- and high-school students. The best system we tried achieved a DER of around 0.34 on our test set and it did not differ statistically significantly across people of different skin tones or genders. Moreover, the system was able to estimate the proportion of speech by different speakers in the group with a correlation of up to 0.62 compared to human annotations. Additionally, it was also capable of distinguishing between the speech of teachers and children with an accuracy of 69.17%.

Limitations: The limitations of this study primarily stem from dataset constraints. For the Sensor Immersion Dataset, there are only 32 audio recordings, which significantly restrict the amount of available test data and prevent us from fine-tuning the model with this data. Additionally, the lack of teachers' enrollments forces us to treat all teachers' speech as non-target speech. Furthermore, the mono-channel nature of the audio recordings limits the applicability of location-based speaker identification methods like direction of arrival (DOA) estimation (Araki et al., 2008). Regarding the NCRECE dataset, the absence of speaker enrollments hinders the use of the nearest enrollment approach in the speaker assignment component of the framework.

Future work: Instead of embedding-based diarization systems such as ECAPA-TDNN, we can use end-to-end neural models such as (Zhang et al., 2022; Fujita et al., 2019; He et al., 2022). These models offer the opportunity to capture simultaneous speech from multiple speakers and might achieve a better DER. With a system capable of identifying simultaneous speech, we could also detect automatically when one student *interrupts* another; this could serve as useful feedback for students, helping to ensure that each person's contributions are heard. We also plan to explore multimodal systems (Kang et al., 2020). By incorporating additional video, image, and text information, we aim to provide more useful data for the task as well as offer greater possibilities for downstream applications.

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.

REFERENCES

ALHARBI, W. 2023. AI in the foreign language classroom: A pedagogical overview of automated writing assistance tools. *Education Research International 2023,* 1, 4253331.

AMAZON. 2021. Amazon transcribe.

ANGUERA, X., BOZONNET, S., EVANS, N., FREDOUILLE, C., FRIEDLAND, G., AND VINYALS, O. 2012. Speaker diarization: A review of recent research. *IEEE Transactions on Audio, Speech, and Language Processing 20,* 2, 356–370.

ARAKI, S., FUJIMOTO, M., ISHIZUKA, K., SAWADA, H., AND MAKINO, S. 2008. A DOA based speaker diarization system for real meetings. In *2008 Hands-Free Speech Communication and Microphone Arrays*. IEEE, 29–32.

BAIN, M., HUH, J., HAN, T., AND ZISSERMAN, A. 2023. Whisperx: Time-accurate speech transcription of long-form audio. *arXiv preprint arXiv:2303.00747*.

BAO, W. 2020. Covid-19 and online teaching in higher education: A case study of Peking University. *Human Behavior and Emerging Technologies 2,* 2, 113–115.

BECCARO, W., ARJONA RAMÍREZ, M., LIAW, W., AND GUIMARÃES, H. R. 2024. Analysis of oral exams with speaker diarization and speech emotion recognition: A case study. *IEEE Transactions on Education 67,* 1, 74–86.

BREDIN, H. 2017. pyannote. metrics: A toolkit for reproducible evaluation, diagnostic, and error analysis of speaker diarization systems. In *Proceedings of the Annual Conference of the International Speech Communication Association (INTERSPEECH)*, F. Lacerda, Ed. International Speech Communication Association (ISCA), 3587–3591.

CAO, J., GANESH, A., CAI, J., SOUTHWELL, R., PERKOFF, E. M., REGAN, M., KANN, K., MARTIN, J. H., PALMER, M., AND D'MELLO, S. 2023. 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*. Association for Computing Machinery, New York, NY, USA, 250–262.

CHEN, S., WANG, C., CHEN, Z., WU, Y., LIU, S., CHEN, Z., LI, J., KANDA, N., YOSHIOKA, T., XIAO, X., WU, J., ZHOU, L., REN, S., QIAN, Y., QIAN, Y., WU, J., ZENG, M., YU, X., AND WEI, F. 2022. WavLM: Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing 16,* 6, 1505–1518.

DESPLANQUES, B., THIENPONDT, J., AND DEMUYNCK, K. 2020. ECAPA-TDNN: Emphasized channel attention, propagation and aggregation in tdnn based speaker verification. *arXiv preprint arXiv:2005.07143*.

DUTTA, S., IRVIN, D., BUZHARDT, J., AND HANSEN, J. H. 2022. 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)*. Association for Computational Linguistics, Seattle, Washington, 92–100.

FITZPATRICK, T. B. 1988. The Validity and Practicality of Sun-Reactive Skin Types I Through VI. *Archives of Dermatology 124,* 6 (06), 869–871.

FUJITA, Y., KANDA, N., HORIGUCHI, S., XUE, Y., NAGAMATSU, K., AND WATANABE, S. 2019. End-to-end neural speaker diarization with self-attention. In *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 296–303.

FUNG, D. C.-L., TO, H., AND LEUNG, K. 2016. 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.

GAZAWY, Q., BUYRUKOGLU, S., AND AKBAS, A. 2023. Deep learning for enhanced education quality: Assessing student engagement and emotional states. In *2023 Innovations in Intelligent Systems and Applications Conference (ASYU)*. IEEE, 1–8.

GOMEZ, A., PATTICHIS, M. S., AND CELEDÓN-PATTICHIS, S. 2022. Speaker diarization and identification from single channel classroom audio recordings using virtual microphones. *IEEE Access 10*, 56256–56266.

GOOGLECLOUD. 2021. Detect different speakers in an audio recording.

HAGEN, A., PELLOM, B., AND COLE, R. 2003. Children's speech recognition with application to interactive books and tutors. In *2003 IEEE Workshop on Automatic Speech Recognition and Understanding (IEEE Cat. No.03EX721)*. IEEE, 186–191.

HARRINGTON, J., PALETHORPE, S., WATSON, C. I., ET AL. 2007. Age-related changes in fundamental frequency and formants: a longitudinal study of four speakers. In *Proceedings of the Annual Conference of the International Speech Communication Association (INTERSPEECH)*. International Speech Communication Association (ISCA), 2753–2756.

HAZIRBAS, C., BITTON, J., DOLHANSKY, B., PAN, J., GORDO, A., AND FERRER, C. C. 2022. Towards measuring fairness in AI: The casual conversations dataset. *IEEE Transactions on Biometrics, Behavior, and Identity Science 4,* 3, 324–332.

HE, M.-K., DU, J., AND LEE, C.-H. 2022. End-to-end audio-visual neural speaker diarization. In *Proceedings of the Annual Conference of the International Speech Communication Association (INTERSPEECH)*. International Speech Communication Association, 1461–1465.

HE, X., WANG, J., TRINH, V. A., MCREYNOLDS, A., AND WHITEHILL, J. 2024. Tracking classroom movement patterns with person re-ID. In *Proceedings of the 17th International Conference on Educational Data Mining*, B. PaaÃŸen and C. D. Epp, Eds. International Educational Data Mining Society, 679–685.

HOWARD, J. R. 2015. *Discussion in the college classroom: Getting your students engaged and participating in person and online*. John Wiley & Sons, San Francisco.

KANG, W., ROY, B. C., AND CHOW, W. 2020. Multimodal speaker diarization of real-world meetings using d-vectors with spatial features. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 6509–6513.

KELLY, S., OLNEY, A. M., DONNELLY, P., NYSTRAND, M., AND D'MELLO, S. K. 2018. Automatically measuring question authenticity in real-world classrooms. *Educational Researcher 47,* 7, 451–464.

LANDINI, F., PROFANT, J., DIEZ, M., AND BURGET, L. 2022. Bayesian HMM clustering of x-vector sequences (VBx) in speaker diarization: theory, implementation and analysis on standard tasks. *Computer Speech Language 71*, 101254.

LUCZYNSKI, K. C. AND HANLEY, G. P. 2013. Prevention of problem behavior by teaching functional communication and self-control skills to preschoolers. *Journal of Applied Behavior Analysis 46,* 2, 355–368.

MAHMOOD, S. 2021. Instructional strategies for online teaching in covid-19 pandemic. *Human Behavior and Emerging Technologies 3,* 1, 199–203.

MARKOV, K. AND NAKAMURA, S. 2008. Improved novelty detection for online GMM based speaker diarization. In *Proceedings of the Annual Conference of the International Speech Communication Association (INTERSPEECH)*. International Speech Communication Association, 363–366.

OLNEY, A. M., DONNELLY, P. J., SAMEI, B., AND D'MELLO, S. K. 2017. Assessing the dialogic properties of classroom discourse: Proportion models for imbalanced classes. In *Proceedings of the 10th International Conference on Educational Data Mining*, X. Hu, T. Barnes, A. Hershkovitz, and L. Paquette, Eds. International Educational Data Mining Society, 162–167.

PARK, T. J., KANDA, N., DIMITRIADIS, D., HAN, K. J., WATANABE, S., AND NARAYANAN, S. 2022. A review of speaker diarization: Recent advances with deep learning. *Computer Speech Language 72*, 101317.

PIANTA, R. AND BURCHINAL, M. 2016. National center for research on early childhood education teacher professional development study (2007-2011). https://doi.org/10.3886/ICPSR34848.v2.

PIANTA, R., HAMRE, B., DOWNER, J., BURCHINAL, M., WILLIFORD, A., LOCASALE-CROUCH, J., HOWES, C., LA PARO, K., AND SCOTT-LITTLE, C. 2017. Early childhood professional development: Coaching and coursework effects on indicators of children's school readiness. *Early Education and Development 28,* 8, 956–975.

PRADHAN, S. S., COLE, R. A., AND WARD, W. H. 2023. My science tutor (myst)–a large corpus of children's conversational speech. In *International Conference on Language Resources and Evaluation*.

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

RADFORD, A., KIM, J. W., XU, T., BROCKMAN, G., MCLEAVEY, C., AND SUTSKEVER, I. 2023. Robust speech recognition via large-scale weak supervision. In *International Conference on Machine Learning*. PMLR, JMLR.org, 28492–28518.

RAMAKRISHNAN, A., ZYLICH, B., OTTMAR, E., LOCASALE-CROUCH, J., AND WHITEHILL, J. 2023. Toward automated classroom observation: Multimodal machine learning to estimate class positive climate and negative climate. *IEEE Transactions on Affective Computing 14,* 1, 664–679.

RAVANELLI, M., PARCOLLET, T., MOUMEN, A., DE LANGEN, S., SUBAKAN, C., PLANTINGA, P., WANG, Y., MOUSAVI, P., LIBERA, L. D., PLOUJNIKOV, A., PAISSAN, F., BORRA, D., ZAIEM, S., ZHAO, Z., ZHANG, S., KARAKASIDIS, G., YEH, S.-L., CHAMPION, P., ROUHE, A., BRAUN, R., MAI, F., ZULUAGA-GOMEZ, J., MOUSAVI, S. M., NAUTSCH, A., LIU, X., SAGAR, S., DURET, J., MDHAFFAR, S., LAPERRIERE, G., ROUVIER, M., MORI, R. D., AND ESTEVE, Y. 2024. Open-source conversational AI with SpeechBrain 1.0.

RAVANELLI, M., PARCOLLET, T., PLANTINGA, P., ROUHE, A., CORNELL, S., LUGOSCH, L., SUBAKAN, C., DAWALATABAD, N., HEBA, A., ZHONG, J., CHOU, J.-C., YEH, S.-L., FU, S.-W., LIAO, C.-F., RASTORGUEVA, E., GRONDIN, F., ARIS, W., NA, H., GAO, Y., MORI, R. D., AND BENGIO, Y. 2021. Speechbrain: A general-purpose speech toolkit.

REYNOLDS, D. A., QUATIERI, T. F., AND DUNN, R. B. 2000. Speaker verification using adapted gaussian mixture models. *Digital Signal Processing 10,* 1-3, 19–41.

ROUVIER, M., BOUSQUET, P.-M., AND FAVRE, B. 2015. Speaker diarization through speaker embeddings. In *2015 23rd European Signal Processing Conference (EUSIPCO)*. IEEE, 2082–2086.

SEDOVA, K., SEDLACEK, M., SVARICEK, R., MAJCIK, M., NAVRATILOVA, J., DREXLEROVA, A., KYCHLER, J., AND SALAMOUNOVA, Z. 2019. Do those who talk more learn more? the relationship between student classroom talk and student achievement. *Learning and Instruction 63*, 101217.

SHOBAKI, KHALDOUN, HOSOM, JOHN-PAUL, AND COLE, RONALD ALLAN. 2007. CSLU: Kids' Speech Version 1.1.

SILERO TEAM. 2024. Silero VAD: pre-trained enterprise-grade voice activity detector (VAD), number detector and language classifier. https://github.com/snakers4/silero-vad.

SOUTHWELL, R., PUGH, S. L., PERKOFF, E. M., CLEVENGER, C., BUSH, J. B., LIEBER, R., WARD, W. H., FOLTZ, P. W., AND D'MELLO, S. K. 2022. Challenges and feasibility of automatic speech recognition for modeling student collaborative discourse in classrooms. In *Educational Data Mining*. International Educational Data Mining Society.

SÜMER, Ö., GOLDBERG, P., D'MELLO, S., GERJETS, P., TRAUTWEIN, U., AND KASNECI, E. 2021. Multimodal engagement analysis from facial videos in the classroom. *IEEE Transactions on Affective Computing 14,* 2, 1012–1027.

THOMAS, D. R., LIN, J., BHUSHAN, S., ABBOUD, R., GATZ, E., GUPTA, S., AND KOEDINGER, K. R. 2024. Learning and ai evaluation of tutors responding to students engaging in negative self-talk. In *Proceedings of the Eleventh ACM Conference on Learning @ Scale*. L@S '24. Association for Computing Machinery, New York, NY, USA, 481–485.

WANG, J., DUDY, S., HE, X., WANG, Z., SOUTHWELL, R., AND WHITEHILL, J. 2024. Speaker diarization in the classroom: How much does each student speak in group discussions? In *Proceedings of the 17th International Conference on Educational Data Mining*, B. PaaÃŸen and C. D. Epp, Eds. International Educational Data Mining Society, 360–367.

WHITEHILL, J. AND LOCASALE-CROUCH, J. 2023. Automated evaluation of classroom instructional support with llms and bows: Connecting global predictions to specific feedback. *Journal of Educational Data Mining 16,* 1, 34–60.

YANG, Q., ZIMMERMANN, K., BARTHOLOMEW, C. P., PURTELL, K. M., AND ANSARI, A. 2023. Preschool classroom age composition and physical literacy environment: Influence on children's emergent literacy outcomes. *Early Education and Development 35*, 1–18.

ZHANG, C., SHI, J., WENG, C., YU, M., AND YU, D. 2022. 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)*. IEEE, 8372–8376.

APPENDIX

Prediction Level	Aggregation Method	Embedding Model	VAD Method				Speech Enhancement	Enrollment Type	Nearest Enrollment VS. Clustering	DER	PCC	SCC
			Whisper	CRDNN (threshold)	Non-target Speech Enrollment	Silero (threshold)						
Frame-wise	/	Fine-tuned ECAPA-TDNN	√	0.8798	√	/	/	entire	Clustering_K-means	0.3412	0.4497	0.5171
Sentence-wise	Sentence Embedding	Fine-tuned ECAPA-TDNN	√	0.8798	√	/	/	entire	Clustering_K-means	0.3446	0.5516	0.6208
Sentence-wise	Argmax Vote	Fine-tuned ECAPA-TDNN	√	0.8798	√	/	/	entire	Nearest Enrollment	0.3508	0.4887	0.5856
Sentence-wise	Weighted Vote	Fine-tuned ECAPA-TDNN	√	0.8798	√	/	/	entire	Nearest Enrollment	0.3508	0.4887	0.5856
Frame-wise	/	Fine-tuned ECAPA-TDNN	√	0.8798	√	/	/	entire	Nearest Enrollment	0.3508	0.4887	0.5856
Sentence-wise	Majority Vote	Fine-tuned ECAPA-TDNN	√	0.8798	√	/	/	entire	Nearest Enrollment	0.3520	0.4927	0.6370
Sentence-wise	Sentence Embedding	Fine-tuned ECAPA-TDNN	√	0.9	√	/	/	entire	Nearest Enrollment	0.3566	0.5045	0.6183
Sentence-wise	Sentence Embedding	Fine-tuned ECAPA-TDNN	√	0.8798	√	/	/	entire	Nearest Enrollment	0.3577	0.5102	0.6165
Sentence-wise	Sentence Embedding	Fine-tuned ECAPA-TDNN	√	0.95	√	/	/	entire	Nearest Enrollment	0.3593	0.4176	0.4874
Sentence-wise	Sentence Embedding	Fine-tuned ECAPA-TDNN	√	/	√	0.8	/	entire	Nearest Enrollment	0.3602	0.4564	0.6190
Sentence-wise	Sentence Embedding	Pre-trained ECAPA-TDNN	√	/	√	0.9	/	entire	Nearest Enrollment	0.3689	0.4494	0.4672
Sentence-wise	Sentence Embedding	Pre-trained ECAPA-TDNN	√	0.95	√	/	/	entire	Nearest Enrollment	0.3699	0.4190	0.5274
Sentence-wise	Sentence Embedding	Pre-trained ECAPA-TDNN	√	0.8798	√	/	/	chunk(4sec)	Nearest Enrollment	0.3745	0.4451	0.4489
Sentence-wise	Sentence Embedding	Fine-tuned ECAPA-TDNN	/	/	√	0.95	/	entire	Nearest Enrollment	0.3746	0.6139	0.6848
Sentence-wise	Sentence-embedding	Fine-tuned ECAPA-TDNN	√	0.8798	√	/	/	entire	Clustering_Agglomerative	0.3746	0.3084	0.4875
Sentence-wise	Sentence Embedding	Pre-trained ECAPA-TDNN	√	0.8798	√	/	/	chunk(8sec)	Nearest Enrollment	0.3751	0.5244	0.5613
Sentence-wise	Sentence Embedding	Pre-trained ECAPA-TDNN	√	0.8798	√	/	/	chunk(2sec)	Nearest Enrollment	0.3772	0.3302	0.3663
Sentence-wise	Sentence Embedding	Pre-trained ECAPA-TDNN	√	0.8798	√	/	/	entire	Clustering_K-means	0.3796	0.5429	0.5843
Sentence-wise	Sentence Embedding	Pre-trained ECAPA-TDNN	√	0.8798	√	/	/	chunk(16sec)	Nearest Enrollment	0.3798	0.5271	0.5521
Sentence-wise	Majority Vote	Pre-trained ECAPA-TDNN	√	0.8798	√	/	/	entire	Nearest Enrollment	0.3838	0.4940	0.5303
Sentence-wise	Argmax Vote	Pre-trained ECAPA-TDNN	√	0.8798	√	/	/	entire	Nearest Enrollment	0.3838	0.4940	0.5303
Sentence-wise	Weighted Vote	Pre-trained ECAPA-TDNN	√	0.8798	√	/	/	entire	Nearest Enrollment	0.3838	0.4940	0.5303
Frame-wise	/	Pre-trained ECAPA-TDNN	√	0.8798	√	/	/	entire	Nearest Enrollment	0.3838	0.4940	0.5303
Frame-wise	/	Pre-trained ECAPA-TDNN	√	0.8798	√	/	/	entire	Clustering_K-means	0.3845	0.2954	0.3540
Sentence-wise	Sentence Embedding	Pre-trained ECAPA-TDNN	√	0.8798	√	/	/	chunk(32sec)	Nearest Enrollment	0.3863	0.4932	0.5474
Frame-wise	/	Fine-tuned ECAPA-TDNN	/	/	√	/	/	entire	NearestEnrollment	0.3867	0.5323	0.6180
Frame-wise	/	Fine-tuned ECAPA-TDNN	/	0.8798	√	/	/	entire	NearestEnrollment	0.3867	0.5323	0.6180
Sentence-wise	Sentence-embedding	Fine-tuned ECAPA-TDNN	√	/	√	/	/	entire	Clustering_K-means	0.3867	0.4345	0.5669
Sentence-wise	Sentence Embedding	Pre-trained ECAPA-TDNN	√	0.9	√	/	/	entire	Nearest Enrollment	0.3876	0.5203	0.6085
Sentence-wise	Sentence Embedding	Pre-trained ECAPA-TDNN	√	0.8798	√	/	/	entire	Nearest Enrollment	0.3937	0.5343	0.6134
Sentence-wise	Sentence Embedding	Pre-trained ECAPA-TDNN	/	/	√	0.95	/	entire	Nearest Enrollment	0.4065	0.6416	0.6411
Sentence-wise	Sentence Embedding	Pre-trained WavLM	√	0.8798	√	/	/	entire	Nearest Enrollment	0.4132	0.2476	0.2912
Sentence-wise	Sentence Embedding	Pre-trained ECAPA-TDNN	√	0.95	/	/	/	entire	Nearest Enrollment	0.4134	0.3190	0.3771
Sentence-wise	Sentence Embedding	Fine-tuned ECAPA-TDNN	√	0.95	/	/	/	entire	Nearest Enrollment	0.4156	0.1949	0.3255
Sentence-wise	Sentence Embedding	Pre-trained ECAPA-TDNN	√	0.8798	√	/	√	entire	Nearest Enrollment	0.4262	0.3723	0.4282
Sentence-wise	Sentence Embedding	Pre-trained ECAPA-TDNN	√	0.9	/	/	/	entire	Nearest Enrollment	0.4648	0.3983	0.4554
Sentence-wise	Sentence Embedding	Fine-tuned ECAPA-TDNN	√	0.9	/	/	/	entire	Nearest Enrollment	0.4676	0.2744	0.3589
Sentence-wise	Sentence Embedding	Pre-trained ECAPA-TDNN	√	0.8798	/	/	/	entire	Nearest Enrollment	0.4792	0.3857	0.4306
Sentence-wise	Sentence Embedding	Fine-tuned ECAPA-TDNN	√	0.8798	/	/	/	entire	Nearest Enrollment	0.4811	0.2543	0.3367
Frame-wise	/	Pre-trained ECAPA-TDNN	/	/	√	/	/	entire	NearestEnrollment	0.5252	0.4991	0.5343
Sentence-wise	Sentence Embedding	Pre-trained ECAPA-TDNN	/	(0.98, 0.8)	√	/	/	entire	Nearest Enrollment	0.5384	0.3996	0.3976
Frame-wise	/	Pre-trained ECAPA-TDNN	/	0.8798	/	/	/	entire	Nearest Enrollment	0.5681	0.1263	0.0705
Sentence-wise	Sentence Embedding	Pre-trained ECAPA-TDNN	√	/	√	/	/	entire	Nearest Enrollment	0.5708	0.5317	0.5518
Sentence-wise	Sentence Embedding	Pre-trained ECAPA-TDNN	/	(0.5, 0.25)	√	/	/	entire	Nearest Enrollment	0.7248	0.3344	0.3280
Sentence-wise	Sentence-embedding	Pre-trained ECAPA-TDNN	√	/	/	/	/	entire	NearestEnrollment	0.7361	0.3142	0.3263
Sentence-wise	Sentence-embedding	Fine-tuned ECAPA-TDNN	√	/	/	/	/	entire	Clustering_K-means	0.7505	-0.0415	0.0610
Frame-wise	/	Pre-trained ECAPA-TDNN	/	(0.5, 0.25)	/	/	/	entire	Nearest Enrollment	0.7575	0.2363	0.3091
Sentence-wise	Sentence-embedding	Fine-tuned ECAPA-TDNN	√	0.8798	√	/	/	entire	Clustering_mean_shift	/	/	/
Sentence-wise	Sentence-embedding	Fine-tuned ECAPA-TDNN	√	0.8798	√	/	/	entire	Clustering_DBSCAN	/	/	/

Figure 4: All experimental results