Estimating Classroom Engagement through Audio-Based Signal Processing and Interaction Analysis

Aneesh Gupta[1], Deepak Yadav[2], Prikshit Juneja[3], Prof. Shampa Chakraverty[4]
[1,2,3]*UG Student, Dept. of CSE, Netaji Subhas University of Technology, New Delhi*
[4]*Professor, Dept. of CSE, Netaji Subhas University of Technology, New Delhi*

***Abstract*—Classroom engagement is an essential element of good pedagogy, however it continues to pose difficulties in objective measurement and analysis. This study shows a new way to measure student involvement in the classroom using audio input. The goal is to give the teachers useful information they can use to improve the way they teach. The suggested method looks at audio from classrooms to create a "interactivity score," which gives a full picture of how teachers and students talk to each other. This analysis makes thorough summaries that include the number of interactions, the tone of the conversations, the topics that were talked about, who was talking, and how long and how important the conversations were. Beyond conventional classrooms, this research finds value in many fields including university and school teaching, army training camps, faculty development programs (FDPs), Corporate Training Sessions, Online Learning Platforms etc. These realizations can enable teachers to improve their instructional strategies so guaranteeing efficient knowledge transmit and skill development.**

This research employs sophisticated signal processing methodologies, encompassing spectral analysis via Mel-frequency cepstral coefficients (MFCCs), clustering algorithms for speaker recognition, and natural language processing (NLP) for sentiment and topic evaluation. The results emphasize the promise of audio-based engagement analysis as a revolutionary instrument in contemporary education and training, facilitating data-driven enhancements in teaching and learning experiences.

***Index Terms*—audio analysis, classroom dynamics, classroom interactivity, education enhancement, educational technology, engagement, engagement detection, interactivity analysis, mel-frequency cepstral coefficients, speaker diarization**

I. INTRODUCTION

Classroom engagement is crucial for effective learning and academic success. Traditional methods like manual observations, surveys, and self-reports are subjective, time-consuming, and difficult to scale for large classrooms. These methods introduce bias and inconsistency, making it challenging for educators to adapt their teaching strategies dynamically based on student responses. Advancements in audio signal processing and artificial intelligence offer an opportunity to develop automated methods for measuring classroom engagement. By leveraging machine learning and speech processing techniques, educators can assess engagement more effectively, providing data-driven insights that help optimize their teaching methods.

This study proposes an audio-based approach to estimate classroom engagement by analysing speech interactions. The system captures and quantifies key metrics such as the *number of questions asked by students*, *teacher responses*, and *diversity of student participation*. Mel-frequency cepstral coefficients (MFCCs) are used to differentiate voices and identify unique speech characteristics. By processing classroom recordings with MFCCs, machine learning models, and natural language processing (NLP) techniques, a scalable and automated system is developed that provides detailed insights into engagement levels. This research aims to bridge the gap between traditional engagement measurement methods and modern AI-powered solutions, demonstrating real-time audio-based analysis revolutionizing classroom interaction tracking. By doing so, it contributes to the ongoing evolution of smart learning environments, ultimately improving student outcomes and making education more interactive, inclusive, and effective.

II. PROBLEM STATEMENT

Classroom engagement is a critical factor in effective pedagogy, yet it remains challenging to measure and analyse objectively. Traditional methods often fail to provide actionable insights into the dynamics of student-teacher interactions. By leveraging advanced signal processing techniques, including MFCCs for spectral analysis, clustering algorithms for speaker recognition, and NLP for sentiment and topic evaluation, the study introduces an "interactivity score" to quantify classroom dynamics. The goal is to provide educators with comprehensive insights into interaction patterns and identify opportunities for improvement. This involves development of a comprehensive framework to evaluate and report engagement metrics such as - Interactivity Score Calculation, Sentiment Analysis, Context and Topic Detection, Participant Analysis, Personalized Dashboard (A web application that provides a user-friendly, personalized dashboard for educators).

III. LITERATURE SURVEY

This study [1] divides the body of existing literature into seven categories, such as engagement-influencing factors, measurement strategies, and engagement-improving tactics. It draws attention to the importance of behavioural, emotional, and cognitive engagement in promoting academic achievement as well as the growing application of computational techniques to overcome the drawbacks of conventional engagement strategies. This study emphasizes how crucial it is to address multiple facets of engagement in order to improve educational outcomes.

This research [2] investigates machine learning methods for forecasting academic performance (SP) and student engagement (SE) in online learning. It emphasizes that while clustering methods are underutilized despite their potential in SE categorization, classification methods are the most widely used approach, appearing in 88.60% of studies. The review highlights the difficulties in establishing consistent SE levels and stresses how crucial it is to combine explainable AI and feature engineering to increase predictive models' transparency. This work provides insights into computational strategies to improve educational outcomes.

This study [3] presents sophisticated techniques for audio data-based fine-grained classroom activity detection. Utilizing neural network architectures like BiGRU and dilated temporal convolutional models, the study was able to distinguish complex classroom activities like lectures, group projects, and student-teacher interactions with state-of-the-art accuracy. Even under difficult circumstances with unseen instructors and environments, the use of mel-filterbank features, OpenSmile, and PASE+ embeddings improved classification accuracy. This work emphasizes the potential of audio-based signal processing in automating classroom activity analysis.

Before the emergence of automated analysis techniques, classroom engagement was typically assessed through manual observation and surveys. Educators and researchers employed techniques such as *Direct Observations* - Teachers recorded student participation patterns by taking notes on classroom behaviour. *Classroom Assessment Scoring System (CLASS)* - A standardized framework evaluating emotional support and classroom organization.

IV. PROPOSED METHODOLOGY

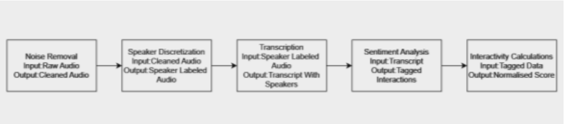


Fig. 4.1 High level implementation

A. Noise Reduction

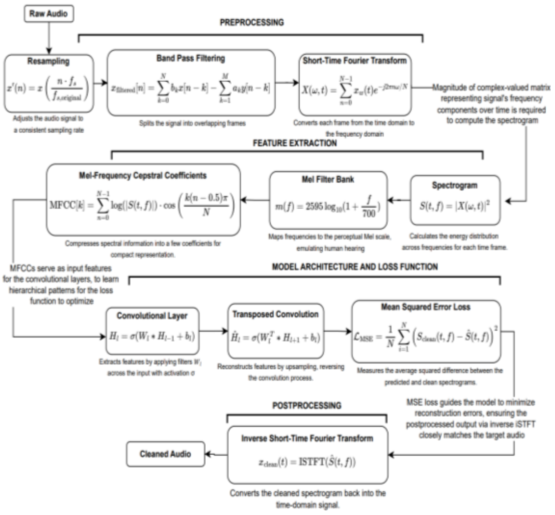

Raw Audio
PREPROCESSING
Resampling
$x'(n) = x\left(\frac{n \cdot f_s}{f_{s,\text{original}}}\right)$
Adjusts the audio signal to a consistent sampling rate
Band Pass Filtering
$x_{\text{filtered}}[n] = \sum_{k=0}^{N} b_k x[n-k] - \sum_{k=1}^{M} a_k y[n-k]$
Splits the signal into overlapping frames
Short-Time Fourier Transform
$X(\omega, t) = \sum_{n=0}^{N-1} x_w(t) e^{-j2\pi n\omega/N}$
Converts each frame from the time domain to the frequency domain
Magnitude of complex-valued matrix representing signal's frequency components over time is required to compute the spectrogram
FEATURE EXTRACTION
Mel-Frequency Cepstral Coefficients
$\text{MFCC}[k] = \sum_{n=0}^{N-1} \log(|S(t,f)|) \cdot \cos\left(\frac{k(n-0.5)\pi}{N}\right)$
Compresses spectral information into a few coefficients for compact representation.
Mel Filter Bank
$m(f) = 2595 \log_{10}(1 + \frac{f}{700})$
Maps frequencies to the perceptual Mel scale, emulating human hearing
Spectrogram
$S(t,f) = |X(\omega,t)|^2$
Calculates the energy distribution across frequencies for each time frame.
MFCCs serve as input features for the convolutional layers, to learn hierarchical patterns for the loss function to optimize
MODEL ARCHITECTURE AND LOSS FUNCTION
Convolutional Layer
$H_l = \sigma(W_l * H_{l-1} + b_l)$
Extracts features by applying filters W_l across the input with activation σ
Transposed Convolution
$\hat{H}_l = \sigma(W_l^T * H_{l+1} + b_l)$
Reconstructs features by upsampling, reversing the convolution process.
Mean Squared Error Loss
$\mathcal{L}_{\text{MSE}} = \frac{1}{N} \sum_{i=1}^{N} \left(S_{\text{clean}}(t,f) - \hat{S}(t,f)\right)^2$
Measures the average squared difference between the predicted and clean spectrograms.
MSE loss guides the model to minimize reconstruction errors, ensuring the postprocessed output via inverse iSTFT closely matches the target audio
POSTPROCESSING
Inverse Short-Time Fourier Transform
$x_{\text{clean}}(t) = \text{ISTFT}(\hat{S}(t,f))$
Converts the cleaned spectrogram back into the time-domain signal.
Cleaned Audio

Fig. 4.2 Noise reduction flow

Noise reduction process is a critical step in preparing audio data for downstream tasks, ensuring the input is clean and devoid of irrelevant or disruptive noise. Begins with preprocessing, involving resampling the audio to a standard sampling rate for uniformity and applying a band-pass filter to isolate the frequency range of interest, effectively removing low - frequency hums and high - frequency artifacts. Next, the audio is divided into short, manageable segments through framing, followed by windowing (typically using a Hamming window) to minimize spectral leakage during transformations. The Short-Time Fourier Transform is then performed on these frames to convert the time-domain signal into its frequency-domain representation, producing a time-frequency spectrogram. This representation facilitates the identification of noise patterns and speech components. Post-STFT, spectral subtraction is applied to suppress noise components while preserving speech signals. Finally, the processed spectrogram undergoes an inverse STFT (iSTFT) to reconstruct the cleaned audio signal back into the time domain.

B. Speaker Diarization

Fig. 4.3 shows the process of dividing audio into smaller parts and identifying speaker parts which involves breaking the audio into frames to preserve important information. From these frames, MFCCs are extracted, capturing unique sound patterns like pitch and tone. A Mel filter bank is used to focus on sounds humans hear best, while a discrete cosine transform reduces complexity. These features are then used to create speaker embeddings, which are like unique fingerprints for each speaker's voice. These embeddings are grouped using agglomerative clustering, combining similar segments until all parts are grouped by speaker. The process stops based on a predefined rule, such as the expected number of speakers. The audio is then divided into segments, labelled according to the speaker, making it easier to analyse or transcribe the audio.

The transition from speaker-labelled audio segments to speaker-labelled transcriptions involves converting speech into text while maintaining the speaker's identity for each segment. First, each speaker-labelled audio segment is passed through an automatic speech recognition system, which processes the audio and converts it into corresponding text. This ensures that spoken content is accurately transcribed for each speaker. The transcribed text is then aligned with the original audio using timestamp matching, ensuring that the transcription remains synchronized with the speech. A contextual language model, such as BERT, is then applied to refine the transcriptions by correcting grammar and ensuring the text is contextually accurate and meaningful as represented in Fig 4.4. Finally, a spelling correction tool, SymSpell, is used to fix any typographical errors, ensuring the text is polished and error-free. The result is a clean, speaker-labelled transcription (for e.g., in Fig 4.5), where each text segment is accurately attributed to the correct speaker, preserving the flow and clarity of the conversation.

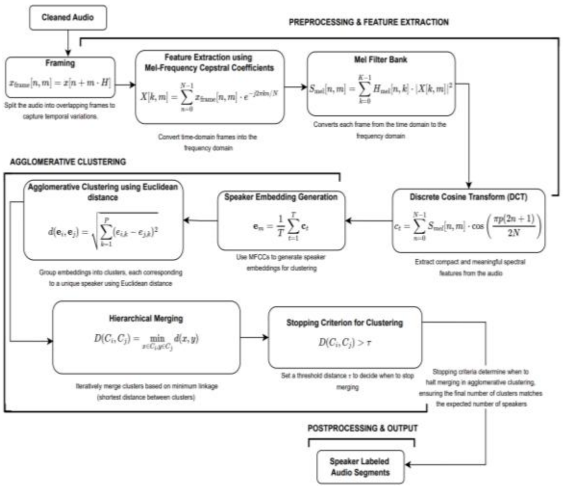

Cleaned Audio
PREPROCESSING & FEATURE EXTRACTION
Framing
$x_{\text{frame}}[n,m] = x[n+m\cdot H]$
Split the audio into overlapping frames to capture temporal variations.
Feature Extraction using Mel-Frequency Cepstral Coefficients
$X[k,m] = \sum_{n=0}^{N-1} x_{\text{frame}}[n,m]\cdot e^{-j2\pi kn/N}$
Convert time-domain frames into the frequency domain
Mel Filter Bank
$S_{\text{mel}}[n,m] = \sum_{k=0}^{K-1} H_{\text{mel}}[n,k]\cdot|X[k,m]|^2$
Converts each frame from the time domain to the frequency domain
AGGLOMERATIVE CLUSTERING
Agglomerative Clustering using Euclidean distance
$d(\mathbf{e}_i,\mathbf{e}_j) = \sqrt{\sum_{k=1}^{P}(e_{i,k}-e_{j,k})^2}$
Group embeddings into clusters, each corresponding to a unique speaker using Euclidean distance
Speaker Embedding Generation
$\mathbf{e}_m = \frac{1}{T}\sum_{t=1}^{T}\mathbf{c}_t$
Use MFCCs to generate speaker embeddings for clustering
Discrete Cosine Transform (DCT)
$c_t = \sum_{n=0}^{N-1} S_{\text{mel}}[n,m]\cdot\cos\left(\frac{\pi p(2n+1)}{2N}\right)$
Extract compact and meaningful spectral features from the audio
Hierarchical Merging
$D(C_i,C_j) = \min_{x\in C_i, y\in C_j} d(x,y)$
Iteratively merge clusters based on minimum linkage (shortest distance between clusters)
Stopping Criterion for Clustering
$D(C_i,C_j) > \tau$
Set a threshold distance τ to decide when to stop merging
Stopping criteria determine when to halt merging in agglomerative clustering, ensuring the final number of clusters matches the expected number of speakers
POSTPROCESSING & OUTPUT
Speaker Labeled Audio Segments

Fig. 4.3 Extract speaker labelled audio segments from cleaned audio

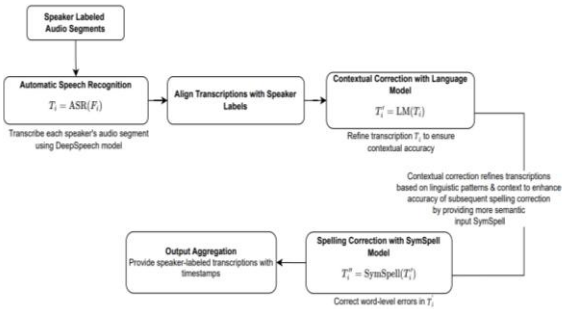
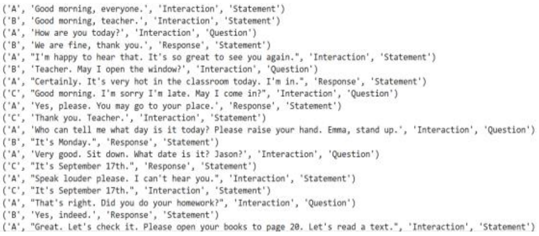

Speaker Labeled Audio Segments
Automatic Speech Recognition
$T_i = \text{ASR}(F_i)$
Transcribe each speaker's audio segment using DeepSpeech model
Align Transcriptions with Speaker Labels
Contextual Correction with Language Model
$T_i' = \text{LM}(T_i)$
Refine transcription T_i to ensure contextual accuracy
Contextual correction refines transcriptions based on linguistic patterns & context to enhance accuracy of subsequent spelling correction by providing more semantic input SymSpell
Output Aggregation
Provide speaker-labeled transcriptions with timestamps
Spelling Correction with SymSpell Model
$T_i'' = \text{SymSpell}(T_i')$
Correct word-level errors in T_i'

Fig. 4.4 From speaker-labelled audio segments to speaker-labelled transcriptions

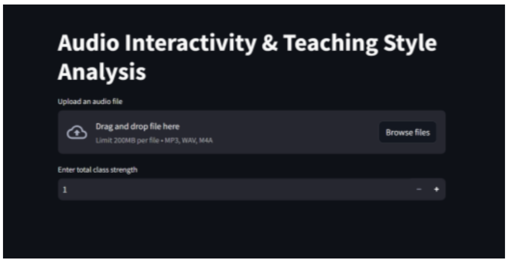


```
('A', 'Good morning, everyone.', 'Interaction', 'Statement')
('B', 'Good morning, teacher.', 'Interaction', 'Statement')
('A', 'How are you today?', 'Interaction', 'Question')
('B', 'We are fine, thank you.', 'Response', 'Statement')
('A', "I'm happy to hear that. It's so great to see you again.", 'Interaction', 'Statement')
('B', 'Teacher. May I open the window?', 'Interaction', 'Question')
('A', "Certainly. It's very hot in the classroom today. I'm in.", 'Response', 'Statement')
('C', "Good morning. I'm sorry I'm late. May I come in?", 'Interaction', 'Question')
('A', 'Yes, please. You may go to your place.', 'Response', 'Statement')
('C', 'Thank you. Teacher.', 'Interaction', 'Statement')
('A', 'Who can tell me what day is it today? Please raise your hand. Emma, stand up.', 'Interaction', 'Question')
('B', "It's Monday.", 'Response', 'Statement')
('A', 'Very good. Sit down. What date is it? Jason?', 'Interaction', 'Question')
('C', "It's September 17th.", 'Response', 'Statement')
('A', "Speak louder please. I can't hear you.", 'Interaction', 'Statement')
('C', "It's September 17th.", 'Interaction', 'Statement')
('A', "That's right. Did you do your homework?", 'Interaction', 'Question')
('B', 'Yes, indeed.', 'Response', 'Statement')
('A', "Great. Let's check it. Please open your books to page 20. Let's read a text.", 'Interaction', 'Statement')
```

Fig. 4.5 Sample output from speaker diarization and transcription analysis

C. Engaging Assessment

The research proposes a web application designed for the educators which comes with a personalized dashboard providing detailed analysis of the uploaded audio.



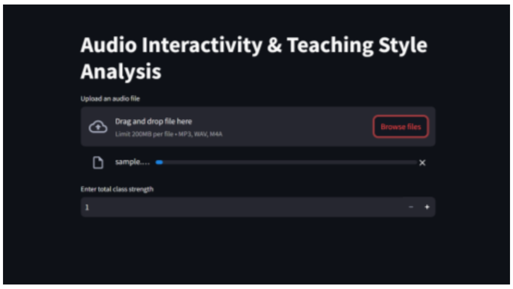
Audio Interactivity & Teaching Style Analysis

Upload an audio file

Drag and drop file here
Limit 200MB per file • MP3, WAV, M4A

Enter total class strength
1

Fig. 4.6 Initial interface of the app



Audio Interactivity & Teaching Style Analysis

Upload an audio file

Drag and drop file here
Limit 200MB per file • MP3, WAV, M4A

Browse files

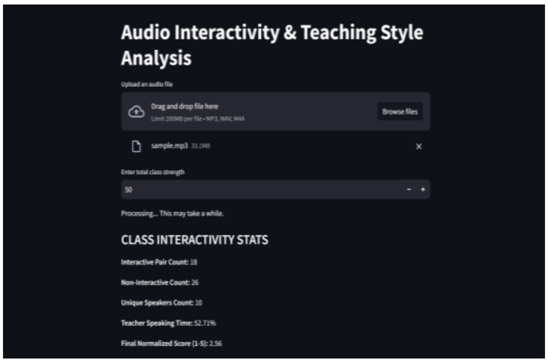
sample....

Enter total class strength

1

Fig. 4.7 Audio recording uploaded in the app

A. Class Interactivity Stats



Audio Interactivity & Teaching Style Analysis

Upload an audio file

Drag and drop file here
Limit 200MB per file • MP3, WAV, M4A

Browse files

sample.mp3 33.1MB

Enter total class strength

50

Processing... This may take a while.

CLASS INTERACTIVITY STATS

Interactive Pair Count: 18

Non-Interactive Count: 26

Unique Speakers Count: 10

Teacher Speaking Time: 52.71%

Final Normalized Score (1-5): 2.56

Fig. 4.8 Class interactivity stats

To calculate the interactivity score, each line of dialogue is evaluated to determine if it represents a question, a response or a general interaction. This is achieved by processing the conversation using a classification function (Fig. 4.10) which examines the sequence of speaker turns and identifies the nature of each utterance based on its content. Questions are detected by checking for the presence of a question mark (?) or relevant words such as "what", "who" etc. in the text. If a question is followed by a response from another speaker, it contributes to the *interactive_pair count*. Interactions between different speakers that are neither questions nor responses contribute to the *non_interactive count*. Using these counts, the interactivity score is calculated with the formula where *unique_speakers* represent number of unique speakers in the interaction and *total_strength* is the total number of participants in the class–

$$value = \frac{(interactive\_pair \times 2)}{(interactive\_pair * 2 + non\_interactive)} + \frac{unique\_speakers}{total\_strength} \tag{1}$$

This formula balances the ratio of question-response interactions to overall interactions, emphasizing engagement through dialogue, and adjusts it based on the diversity of speakers participating in the conversation. By integrating both speaker diversity and interaction dynamics, this method provides a comprehensive measure of classroom interactivity. The *score formula ranges between 0 and 2* because each component is designed to measure normalized interactions. To scale the interactivity score to a normalized range of *1 to 5*, the following formula is applied –

$$score_{\text{normalized}} = scale_{\text{min}} + \frac{(score - score_{\text{min}})(scale_{\text{max}} - scale_{\text{min}})}{score_{\text{max}} - score_{\text{min}}} \tag{2}$$

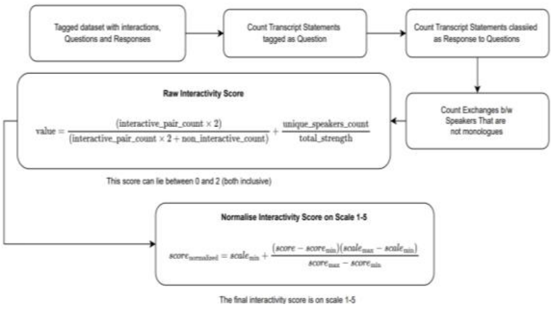

Tagged dataset with interactions, Questions and Responses
Count Transcript Statements tagged as Question
Count Transcript Statements classied as Response to Questions
Count Exchanges b/w Speakers That are not monologues
Raw Interactivity Score
value = (interactive_pair_count × 2) / (interactive_pair_count × 2 + non_interactive_count) + unique_speakers_count / total_strength
This score can lie between 0 and 2 (both inclusive)
Normalise Interactivity Score on Scale 1-5
score_normalized = scale_min + (score − score_min)(scale_max − scale_min) / (score_max − score_min)
The final interactivity score is on scale 1-5

Fig. 4.9 High-level schematic of the interaction score calculation procedure

A crucial component of the formula is the overall class strength which significantly affects the final interactivity score, as it's inversely proportional to the class strength. Increased class sizes diminish the effect of individual interactions, leading to lower overall scores, while smaller class sizes enhance the importance of contributions, promoting higher interactivity metrics.

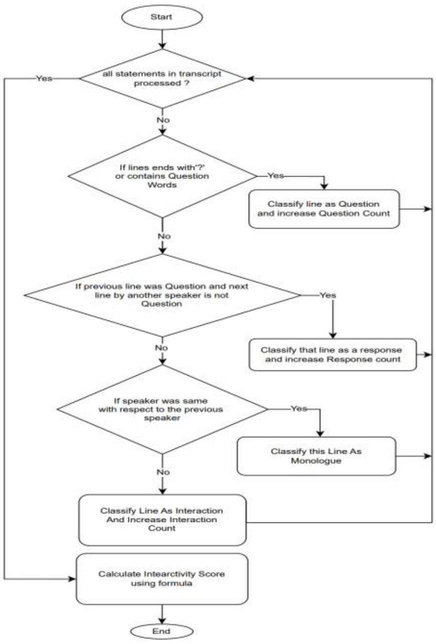

Start
all statements in transcript processed ?
Yes
No
If lines ends with'?' or contains Question Words
Yes
Classify line as Question and increase Question Count
No
If previous line was Question and next line by another speaker is not Question
Yes
Classify that line as a response and increase Response count
No
If speaker was same with respect to the previous speaker
Yes
Classify this Line As Monologue
No
Classify Line As Interaction And Increase Interaction Count
Calculate Intearctivity Score using formula
End

Fig. 4.10 Detailed analytical process involved in the calculation

B. Teaching Style

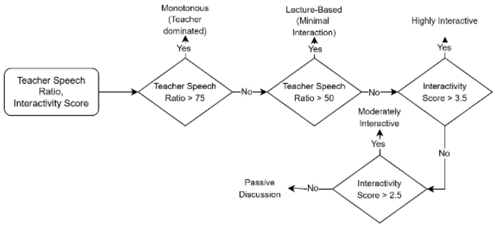

Teacher Speech Ratio, Interactivity Score
Teacher Speech Ratio > 75
Yes
Monotonous (Teacher dominated)
No
Teacher Speech Ratio > 50
Yes
Lecture-Based (Minimal Interaction)
No
Interactivity Score > 3.5
Yes
Highly Interactive
No
Interactivity Score > 2.5
Yes
Moderately Interactive
No
Passive Discussion

Fig. 4.11 Method used for evaluating teaching style

The teaching style in a classroom is categorized based on the ratio of teacher speech to overall speech and the calculated interactivity score. If the teacher's speech accounts for more than 75% of the total audio, the style is classified as *Monotonous (Teacher-Dominated)* indicating minimal student involvement. When the teacher's speech constitutes between 50% to 75%, it's labelled as *Lecture-Based (Moderate Interaction)* reflecting limited engagement from students. For sessions where the interactivity score exceeds 3.5, the style is defined as *Highly Interactive* signifying significant student participation and dynamic discussions. A score between 2.5 and 3.5 results in a *Moderately Interactive* classification, suggesting a balance between teacher-led content and student contributions. Lastly, the style is categorized as *Passive Discussion* when the interactivity score is low, denoting minimal dialogue or collaborative exchanges. This framework provides a structured approach to evaluating teaching effectiveness and engagement levels.

C. Transcription Summary

It's a concise and insightful overview of the classroom session, designed to assist educators in quickly understanding the key points discussed during the lecture. The summary serves as a reference tool, allowing teachers to revisit the main topics and concepts covered without needing to sift through the entire audio.

D. Speaker Activity Timeline

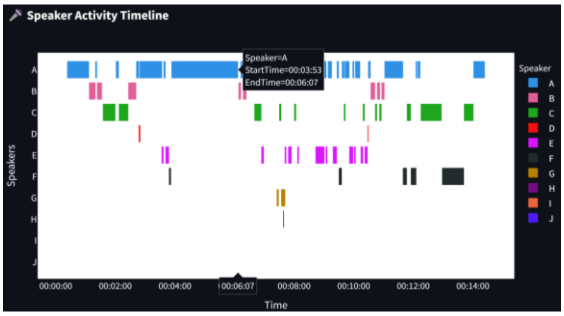


Fig. 4.12 Speaker activity timeline

It's a visual representation that maps the speaking intervals of each participant throughout the session. This feature allows educators to identify the most active participants and understand the flow of interaction between the host (teacher) and attendees (students) as well as participation levels. Additionally, it offers an approximate ratio of teacher-to-student speech, enabling educators to gauge the balance of interaction.

E. Topic Discussion Timeline

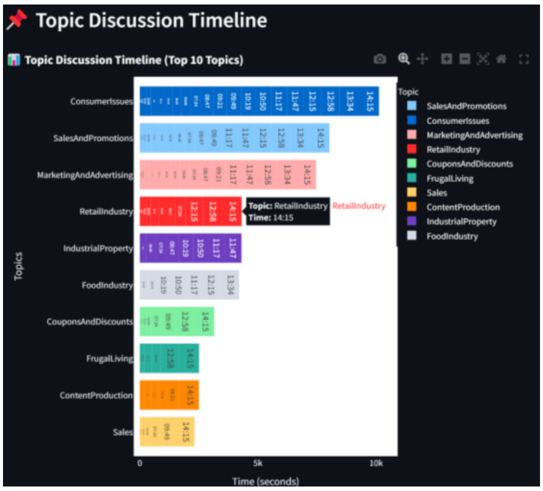


Fig. 4.13 Topic discussion timeline

It's a graphical representation that visualizes the progression of topics discussed during a session, pinpointing the exact time intervals when each subject was addressed. This timeline enables educators to understand the structure and pacing of their lectures, ensuring that time is distributed effectively across various topics. It also highlights the duration spent on each subject, offering insights into the areas that received more attention or required further elaboration. One of the key advantages of this visualization is its ability to help educators identify whether they stayed on track with their planned agenda or deviated to related discussions providing a clear overview of the session, making it easier to prepare follow-up lessons or address any uncovered material

F. Top 5 Topics Discussed in Class

This feature highlights the most frequently covered subjects during a session, providing educators with a concise overview of the core areas of focus. It ranks topics based on their prominence and duration in the discussion. It offers an at-a-glance summary of the session's primary themes, saving time and aiding reflection. For educators, it serves as a valuable tool for assessing the alignment of classroom discussions with lesson objectives and identifying any potential gaps. It can also help in preparing targeted follow-ups or supplementary materials for topics that were heavily emphasized.

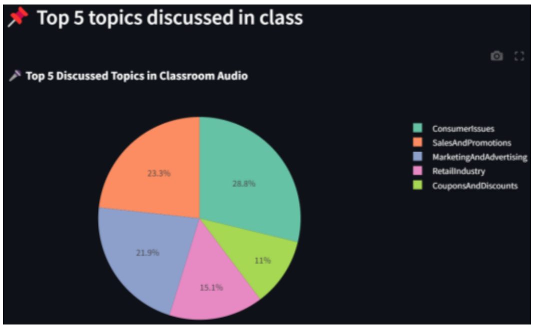


Fig. 4.14 Top 5 topics discussed in class

V. CONCLUSION

The examination of classroom involvement yields a quantitative comprehension of engagement trends derived from auditory data. Utilizing advanced methods like as noise reduction, speaker diarization, and speech-to-text transcription, the procedure provides a thorough framework for assessing interactivity. The extraction of speaker-labelled audio segments facilitates accurate identification of individual contributions, while the incorporation of natural language processing models improves transcription precision and contextual understanding. The interaction score, calculated using a carefully constructed formula, measures the collaborative dynamics of the classroom. The normalization of the score guarantees consistent comparability across different classroom sizes, addressing scalability issues. The categorization of utterances into questions, responses, or interactions elucidates the characteristics of discourse, providing valuable insights about the quality and diversity of involvement.

This methodology establishes a basis for subsequent uses, including adaptive teaching strategies and focused interventions to enhance participation. Future improvements may include real-time processing and integration with video data for a multimodal study, facilitating a comprehensive assessment of classroom interactions.

REFERENCES

[1] Subramainan, Latha & Mahmoud, Moamin. (2020). A Systematic Review on Students' Engagement in Classroom: Indicators, Challenges and Computational Techniques. International Journal of Advanced Computer Science and Applications. 11. 10.14569/IJACSA.2020.0110113.

[2] CHONG, KE & IBRAHIM, NORAINI & Huspi, Sharin Hazlin & Wan Kadir, Wan Mohd Nasir & Isa, Mohd. (2025). A Systematic Review of Machine Learning Techniques for Predicting Student Engagement in Higher Education Online Learning. Journal of Information Technology Education: Research. 24. 005. 10.28945/5456.

[3] Slyman, E., Daw, C., Skrabut, M., Usenko, A., & Hutchinson, B. (2021). Fine-Grained Classroom Activity Detection from Audio with Neural Networks. *arXiv (Cornell University).* https://doi.org/10.48550/arxiv.2107.14369

[4] Falcon, S., Alvarez-Alvarez, C., & Leon, J. (2024). Semi-automated analysis of audio-recorded lessons: The case of teachers' engaging messages. arXiv (Cornell University). https://doi.org/10.48550/arxiv.2412.12062

[5] S. Pradeep Kumar, A. Daripelly, S. M. Rampelli, S. K. R. Nagireddy, A. Badishe and A. Attanthi, "Noise Reduction Algorithm for Speech Enhancement," 2023 International Conference on Signal Processing, Computation, Electronics, Power and Telecommunication (IConSCEPT), Karaikal, India, 2023, pp. 1-5, doi:10.1109/IConSCEPT57958.2023.10170204.