

GuardianEye-AI-Powered Meeting Analyser for Smarter Meetings

Abhishek S
Dept of ISE
Don Bosco Institute of
Technology Bangalore,
India

Mrs. Harshitha HM
Assistant professor Dept of ISE
Don Bosco Institute of
Technology Bangalore, India

Prem Chand
Dept of ISE Don Bosco Institute of
Technology Bangalore,
India

Poojitha M
Dept of ISE Don Bosco Institute of
Technology Bangalore, India

Kotresh R
Dept of ISE Don Bosco Institute of
Technology Bangalore, India

Abstract— GuardianEye AI is a smart meeting application designed to help people understand how a meeting is going. It allows a host to create a meeting and participants to join using a meeting code or link. The application provides a simple and professional meeting interface where users can use their camera, microphone, and other meeting controls.

During or after the meeting, GuardianEye AI studies the participants' communication and behaviour. It provides information about communication skills, fluency, confidence, leadership quality, facial expressions, awkwardness, listening skills, and engagement. It also provides fact-checking information for important statements made during the meeting.

The main purpose of GuardianEye AI is to make meetings more useful by providing a clear report after the meeting. The host can view the performance of each participant in an easy-to-understand table and use the results to identify strengths and areas that need improvement. The system can be useful for online meetings, interviews, presentations, training sessions, and educational discussions.

I. INTRODUCTION

Online meetings have become an important part of education, business, interviews, training, and team discussions. In a normal online meeting, it can be difficult for the host to understand how each participant is communicating, whether they are engaged, and how they are responding during the discussion. Most existing meeting applications mainly provide features such as video calling, audio, screen sharing, chat, and meeting links, but they do not provide a detailed view of the communication and behaviour of participants.

GuardianEye AI is developed as a meeting platform with an additional analysis system for understanding participant performance. The application provides separate interfaces for the host and participants. The host can create and manage meetings, share a meeting code or link, accept or decline participants, and conduct the meeting. Participants can join using the meeting code or link and use basic meeting controls such as camera, microphone, and raise hand. After the meeting, the system provides information about communication skills, communication fluency, confidence, leadership quality, facial expressions, awkwardness, listening skills, and fact-checking.

The main aim of GuardianEye AI is to help the host understand the overall quality of communication during a meeting. The analysis is presented in a simple table and report so that the performance of each participant can be easily compared. The system can be useful for online meetings, interviews, student presentations, training sessions, group discussions, and professional meetings where communication plays an important role.

The system is designed with two different types of users: host and participant. The host has more control over the meeting, such as creating a meeting, sharing the meeting code, managing participants, and starting the analysis. The participant has a simple interface for entering the meeting code, joining the meeting, controlling the camera and microphone, and raising a hand when they want to speak. This separation makes the application easier to use and keeps the meeting process organized.

GuardianEye AI also provides a detailed analysis section after the meeting is completed. The results are arranged in rows and columns, where each participant can be compared based on different communication factors. The host can view the participant's communication skills, fluency, leadership quality, facial expressions, awkwardness, listening ability, confidence, and fact-checking results in one place. This makes it easier to identify areas where a participant is performing well and areas where improvement may be required.

II. RELATED WORK LITERATURE REVIEW

Heo et al. [1] proposed a multimodal emotion recognition system that combines speech, text, and facial information to identify emotions during conversations. Their method detects the actual speaker from video frames by observing facial movements and then uses the speaker's facial expressions along with speech and text for emotion recognition. The system was tested on the MELD dataset and showed better results than the existing methods. This study is useful for our project because it shows how facial expressions, speech, and text can be combined to understand a person's behaviour and emotions during a meeting.

Li and Peng [2] proposed an end-to-end multimodal emotion recognition system that combines facial expressions and remote photoplethysmography (rPPG) signals obtained from facial videos. Their method extracts facial expression features using ResNet50 and processes the rPPG signals using a deep learning network. A Transformer-based Cross-Modal Attention (TCMA) mechanism is then used to learn the relationship between the two types of information. The system was tested on the MAHNOB-HCI dataset and showed better emotion recognition performance when both facial expressions and rPPG signals were combined. This study is useful for our project because it shows how facial information and additional signals can be combined to understand a person's emotional state more effectively during a meeting.

Wenskovitch et al. [3] proposed a framework for understanding interaction uncertainty in human-machine teams, focusing on communication, behaviour, trust, and coordination between humans and intelligent systems. Their method explains different types of uncertainty that can occur during interaction, including situational uncertainty such as recognizing what a teammate is doing, understanding why they are doing it, and predicting their next action. It also considers relational uncertainty, which includes understanding goals, checking whether goals are aligned, and determining how one teammate can influence another. The study discusses how clear communication, visual presentation, and sharing information about decisions can reduce uncertainty and improve trust between team members. This study is useful for our project because it shows how human behaviour, communication, and interaction can be analysed to understand a participant's confidence, engagement, and communication during a meeting.

Zgank and Donaj [4] analyzed how emotional speech affects the performance of an end-to-end automatic speech recognition system. Their study used the Slovenian Artur speech database for training and the Interface emotional speech database for testing different speaking styles and emotions. The researchers considered six emotions, including anger, disgust, fear, joy, sadness, and surprise, along with neutral speech. Their results showed that the speech recognition error increased when the speaker used emotional speech, with anger producing the highest word error rate of 40.02% and surprise producing the lowest at 35.36%. The study also found that changes in speaking speed and emotional style can affect speech recognition accuracy. This study is useful for our project because it shows how speech, emotions, and speaking style can be analyzed together to understand a person's emotional state and communication during a meeting.

Rudresh P et al. [5] proposed a multilingual speaker diarization system to identify and separate different speakers in an audio recording. Their approach divided the audio into speech segments using the PyanNet segmentation model and extracted speaker-specific features using ECAPA-TDNN, X-vector MFCC, and X-vector SincNet models. The extracted speaker embeddings were then grouped using clustering techniques such as agglomerative and spectral clustering. The system was evaluated on a multilingual dataset containing 100 audio recordings, and performance was measured using Diarization Error Rate (DER). Among the compared models, ECAPA-TDNN achieved the lowest DER of 21.3%, followed by X-vector MFCC at 25.0% and X-vector SincNet at 27.9%. This study is useful for our project because it shows how speaker diarization can identify who is speaking and when during a conversation. This can help GuardianEye separate different participants in a meeting and connect their speech with other meeting analysis features such as communication and emotional analysis.

Oleti et al. [6] developed INSIGHTMEET, an AI-powered meeting summarization system that analyzes meeting videos to understand the content, emotions, speakers, and interaction between participants. Their system uses Whisper for speech-to-text transcription, Resemblyzer and Deepgram for speaker diarization, DeepFace for facial emotion analysis, and VADER/Gemini for sentiment analysis. It generates speaker-tagged transcripts, meeting summaries, emotion timelines, participation metrics, and interaction graphs through an interactive dashboard. The system also provides Minutes of Meeting (MOM) and allows the results to be downloaded in formats such as TXT, CSV, and JSON. The study showed that the AI mode achieved faster processing and higher accuracy, with about 95–98% transcription accuracy and around 95% speaker identification accuracy, while the free mode provides privacy-preserving local processing. This study is useful for our project because it shows how speech, speaker identification, emotion analysis, summarization, and interaction patterns can be combined to understand a meeting more effectively. These ideas can help our project analyze meeting discussions, identify participants, understand their emotions and communication patterns, and generate useful meeting summaries.

Hsieh et al. [7] developed an interview recording system that combines speaker recognition, speech-to-text transcription, and speech emotion classification to analyze interview and meeting conversations. Their system uses Microsoft Azure Speech-to-Text API for converting speech into text, while speaker recognition is performed using metric learning and meta-learning with ResNet34 and Triplet Loss. For emotion analysis, a CNN-LSTM-based model with transfer learning is adapted for multi-label emotion classification. The system also includes a Meeting Recording System, Data Labeling System, and Voiceprint Management System for recording meetings, correcting speaker and emotion labels, and managing speaker voiceprints. The speaker recognition model achieved an accuracy of 96.39%, while the multi-label emotion classification achieved an F1-score of 51.15%. This study is useful for our project because it shows how speech transcription, speaker identification, and emotion analysis can be combined to understand meeting conversations. These ideas can help GuardianEye identify participants, analyze their emotions, and generate meaningful meeting records for better meeting analysis.

Rani et al. [8] developed a real-time multimodal emotion detection system that combines visual, audio, and textual information to identify emotions from live video streams. Their system uses EfficientNet for extracting facial and visual features, LSTM for analyzing speech and audio features, and BERT for understanding the text obtained from speech transcription. An attention-based fusion mechanism is used to combine the visual, auditory, and textual features by giving importance to the most useful information from each modality. The system was designed for real-time applications such as virtual meetings, online classes, and live streaming. The proposed model achieved an accuracy of 87.4%, an F1-score of 86.2%, and an average latency of 45 ms per sample. It also achieved an average accuracy of 82.1% under noisy conditions, showing that the system can work effectively in real-world situations. The study is useful for our project because it shows how video, speech, and text can be combined to understand the emotions of people during conversations. These ideas can help GuardianEye analyze participants' facial expressions, voice characteristics, and spoken content together to provide better meeting analysis and meaningful meeting records.

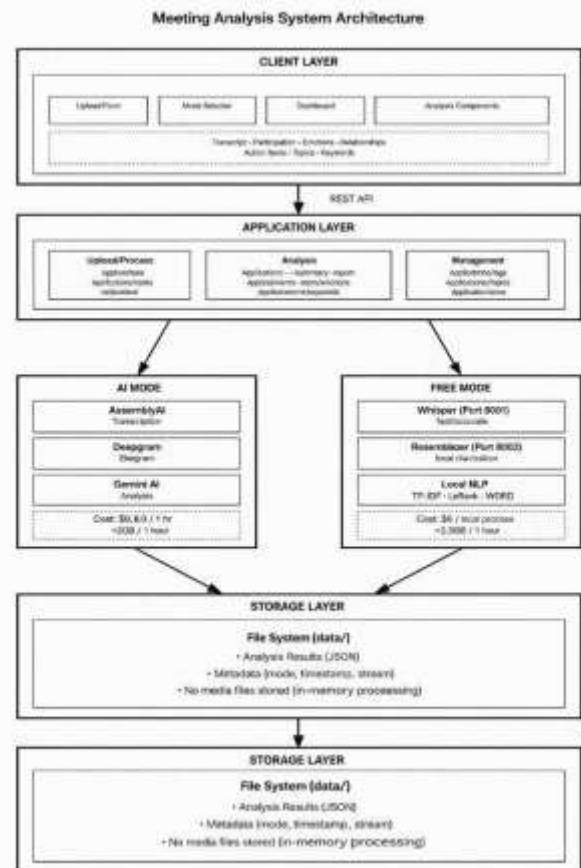
Srivastav et al. [9] developed an automated confidence assessment system that analyzes speech signals to identify the confidence level of a speaker. The system collects speech recordings, performs speech signal preprocessing, extracts important acoustic features such as Mel Frequency Cepstral Coefficients (MFCCs), and uses a Convolutional Neural Network (CNN) to classify the speaker's confidence level. The model categorizes confidence into different levels such as high, medium, and low. The proposed system was trained using labeled speech data and achieved an accuracy of 94%. The study also explains that speech characteristics such as pitch, loudness, speech rate, and intonation can provide useful information about a person's communication and emotional state. This study is useful for our project because it shows how speech signals and extracted audio features can be analyzed to understand the speaker's communication characteristics. These ideas can help GuardianEye analyze participants' voice characteristics during meetings and combine speech-based information with other meeting analysis features to provide better understanding of participants and meaningful meeting records.

Kachhoria et al. [10] developed an automated meeting minutes generation system for online meetings using Natural Language Processing (NLP) and Machine Learning (ML) techniques. The system uses Whisper AI to convert meeting audio into text and pyannote to identify different speakers and assign speaker labels to the transcript. It then uses summarization models such as PEGASUS, BART, LED, Meeting Summary, and bart-conversation-summarizer to generate short and meaningful meeting summaries. The system evaluates speaker identification using Diarization Error Rate (DER) and evaluates the generated summaries using ROUGE metrics. The study achieved a DER between 20% and 35%, while PEGASUS performed best among the tested summarization models with a ROUGE-1 score of 0.35. The study is useful for our project because it shows how speech transcription, speaker identification, and text summarization can be combined to analyze meeting conversations. These ideas can help GuardianEye identify participants, understand their spoken content, and generate meaningful meeting records for better meeting analysis.

Xu et al. [11] proposed a facial expression recognition method based on an Inception Network Module and SENet attention mechanism to improve the recognition of human emotions from facial images. The system uses multi-scale convolution in the Inception module to extract facial features at different levels and introduces the SENet attention mechanism to give more importance to important facial feature channels while reducing unnecessary information. The authors evaluated the proposed model using the FER2013 dataset, which contains 35,887 grayscale facial images belonging to seven emotion categories: Angry, Disgust, Fear, Happy, Sad, Surprise, and Neutral. Data augmentation techniques such as rotation, scaling, translation, and shearing were used to improve the robustness of the model. The proposed network achieved an overall facial expression recognition accuracy of 66.34%, with Happy achieving the highest recognition accuracy of 0.87, while Fear was the most difficult emotion with an accuracy of 0.41. The study is useful for our project because it demonstrates how facial expressions can be analyzed using deep learning and attention mechanisms to identify emotional states. These ideas can help GuardianEye analyze participants' facial expressions during meetings and combine emotion information with other meeting data to provide better understanding of participants' emotions and meeting interactions.

Tabet et al. [12] proposed a deep learning-based student engagement detection method using eye-tracking data to identify whether a person is engaged or not engaged. The proposed system combines the pre-trained VGG16 convolutional neural network with a Long Short-Term Memory (LSTM) network, where VGG16 extracts spatial features from eye images and LSTM captures temporal changes in gaze patterns. The model was evaluated using an Eye Gaze dataset containing approximately 14,500 eye-region images with different gaze directions and engagement labels. The images were resized to 224×224 pixels and data augmentation techniques such as rotation, flipping, and brightness adjustment were applied to improve the robustness of the model. The dataset was divided into 80% training, 10% validation, and 10% testing sets. The proposed VGG16-LSTM model achieved a classification accuracy of 99.57%, performing better than several existing approaches such as CNN-LSTM, CNN-SVM, and Gaze-Swin. The study is useful for our project because it demonstrates how visual information and temporal patterns can be analyzed to understand a participant's engagement and attention. These ideas can help GuardianEye analyze participants' visual engagement during meetings and combine this information with speaker identification, speech transcription, emotion analysis, and meeting summarization to provide a better understanding of meeting interactions and participant behavior.

V. EXISTING SYSTEM FLOW CHART



The InsightMeet system is divided into different layers to make the meeting analysis process simple and organized. The Client Layer provides the user interface, where the user can upload meeting data, select the required mode, view the dashboard, and access analysis results such as transcripts, participation, emotions, relationships, action items, topics, and keywords. The Application Layer receives the user request through the REST API and manages uploading, processing, analysis, and result management. It handles tasks such as generating summaries, identifying emotions, extracting keywords, and managing the meeting data.

The system then provides two processing modes: AI Mode and Free Mode. In AI Mode, AssemblyAI is used for transcription, Deepgram is used for speech-related processing, and Gemini AI is used for meeting analysis. In Free Mode, Whisper performs speech transcription, Resemblyzer is used for speaker diarization, and local NLP techniques such as TF-IDF, TextRank, and Word-based methods are used for text analysis. After processing, the results are stored in the Storage Layer as JSON files along with metadata such as the selected mode, timestamp, and stream information. The system mainly processes the meeting data in memory and does not permanently store the original media files.

VII. Methodology

The proposed GuardianEye AI Meeting Analyzer follows a structured methodology to analyse participant behaviour, communication quality, and overall meeting performance using Artificial Intelligence and Machine Learning techniques. The system is designed to process both live online meetings and pre-recorded meeting videos. The main purpose of the system is to collect meeting data, analyse participant communication and behaviour, and provide useful insights and evaluation results.

The methodology begins with data collection from different meeting sources. The system accepts input either from a live virtual meeting or from an uploaded recorded meeting video. Once the meeting data is received, screen and audio capture processes collect participant video activity and spoken communication. The captured information is then passed to the different analysis modules for further processing.

After collecting the meeting information, participant identification is performed to recognise and separate the individual participants present in the meeting. This helps the system track each participant's interaction, speaking activity, and communication behaviour throughout the meeting. The system can therefore analyse the performance and involvement of individual participants rather than treating the complete meeting as a single conversation.

The captured audio is then converted into text using Speech-to-Text technology. The converted text is processed using Natural Language Processing (NLP) techniques to analyse communication flow, sentence structure, speaking patterns, response relevance, and filler-word usage. NLP analysis helps GuardianEye understand the content of the conversation and identify important communication patterns from the meeting discussion.

At the same time, facial expression and behavioural analysis are performed using computer vision and deep-learning techniques. The system observes participant expressions, eye contact, engagement level, confidence, awkwardness, and emotional behaviour during the meeting. These observations are combined with the communication information to provide a broader understanding of participant behaviour and meeting interaction.

After the communication and behavioural data are analysed, the processed information is transferred to the scoring engine. Machine Learning algorithms such as Decision Tree and Random Forest are used to analyse the collected features and generate participant performance scores. The system evaluates different factors such as communication, participation, engagement, confidence, leadership, behaviour, and other communication patterns. The results are then organized into meaningful performance measures.

Finally, the analysed results are used to generate a meeting summary and result dashboard. The dashboard presents the important findings in an easy-to-understand format, including participant scores, communication insights, behavioural observations, meeting highlights, and overall performance. The system can also provide reports and summaries that help users review the meeting without going through the complete recording again.

VIII. Discussions

The proposed GuardianEye AI Meeting Analyzer provides a more advanced approach to understanding meetings compared with traditional meeting systems. Existing meeting applications mainly focus on recording meetings, converting speech into text, or generating basic summaries. However, these systems may not provide detailed information about participant behaviour, communication quality, confidence, engagement, leadership, and emotional interaction. GuardianEye is designed to provide

The main advantage of the proposed system is that it combines Speech-to-Text, Natural Language Processing (NLP), facial emotion recognition, computer vision, and Machine Learning techniques in a single platform. This combination allows the system to analyse both verbal and visual information during meetings. Speech analysis helps understand what participants say, while facial and behavioural analysis provides information about their expressions, engagement, eye contact, and interaction. Combining these results provides more meaningful participant and meeting insights.

Unlike systems that only generate meeting summaries, GuardianEye also evaluates different aspects of participant performance. The system considers communication skills, participation, engagement, leadership behaviour, confidence, dominance, awkwardness, filler-word usage, and response quality. These results help users understand how participants communicate and interact during a meeting. The analysis can also help identify important discussion points and provide a clearer view of overall meeting performance.

Another important feature of the proposed system is its support for both live online meetings and pre-recorded meeting videos. Users can analyse information from a live meeting or upload an existing recording for analysis. This provides flexibility for different situations, such as regular team meetings, interviews, presentations, and other professional discussions. The stored results and generated reports can also be reviewed later for comparison and further understanding.

The proposed system also provides an analytics dashboard where the analysed information is presented in a simple and organized manner. Users can view participant scores, communication insights, behavioural observations, meeting summaries, and important action items. By bringing these different results together, GuardianEye can reduce the need for users to manually review the complete meeting recording and helps them quickly understand the most important information from the meeting.

Conclusion

The proposed GuardianEye AI Meeting Analyzer provides an intelligent and structured approach to analysing meetings and understanding participant performance. The system combines meeting audio, video, speech-to-text conversion, Natural Language Processing, facial and behavioural analysis, and Machine Learning techniques to study both verbal and non-verbal communication. It supports live online meetings as well as pre-recorded meeting videos, making it useful for different types of professional meetings and discussions.

GuardianEye goes beyond basic meeting recording and summary generation by analysing important factors such as communication quality, participation, engagement, confidence, leadership behaviour, filler words, facial expressions, eye contact, and speaking patterns. Machine Learning algorithms such as Decision Tree and Random Forest are used to support participant evaluation and scoring. The results are presented through an analytics dashboard with meeting summaries, participant insights, performance scores, and reports.

Overall, the proposed system helps users understand meetings more effectively without manually reviewing the entire recording. By combining different types of meeting information in one platform, GuardianEye can provide meaningful communication and behavioural insights. The system can be further improved in the future by using larger datasets, more advanced models, improved emotion recognition, and additional meeting analysis features to provide more accurate and useful results.

References

1. Hsieh, W. Y., Tsai, H. C., Chen, L. A., & Ĩk, T. U. (2024). Developing an Interview Recording System with Speaker Recognition and Emotion Classification.
2. Heo, S., Kyung, J., & Chang, J. H. (2024). Multimodal Emotion Recognition with Target Speaker-Based Facial Embeddings.
3. Li, J., & Peng, J. (2024). End-to-End Multimodal Emotion Recognition Based on Facial Expressions and Remote Photoplethysmography Signals.
4. Khaoula, T., & Soltane, M. (2025). Deep Learning Based Student Engagement Detection Using Eye Tracking.
5. Xu, Y., Yu, Y., & Wu, B. (2025). Facial Expression Recognition Based on Inception Network Module.
6. Kachhoria, R., Daga, N., Ramteke, H., Akotkar, Y., & Ghule, S. (2023). Minutes of Meeting Generation for Online Meetings Using NLP & ML Techniques.
7. Slovenian End-to-End Automatic Speech Recognition and the Analysis of Emotional Speech.
8. Oleti, B. L., Prakaash, R. A. D., RanjithKumar, G. S., Reddy, P. P., Narendatta, N., & Vignesh, C. C. (2025). INSIGHTMEET: An AI Powered Meeting Summarizer with Emotion Analytics and Interaction-Mapping.
9. Srivastav, K., Sharma, S., Srivastava, M., & Singhal, K. (2024). Decoding Communication: Confidence Assessment through CNN-based Speech Signals Analysis.
10. Wenskovitch, J., Fallon, C., Miller, K., & Dasgupta, A. (2024). Characterizing Interaction Uncertainty in Human-Machine Teams.
11. Rani, S., Muralidhar, A., Rao, G. S., Shareef, S. K., Sudheer, P., & Bhavsingh, M. (2024). Real-Time Emotion Detection in Live Video Streams using Multimodal Deep Learning Architectures.
12. Rudresh, P., Tharunn Raj, P. V., Hariharan, K., Sivaprasath, S., Sherin Shibi, C., Jeyasudha, J., Vikram, C. M., Agarwal, G. K., & Gohil, R. P. (2024). Speaker Diarization in Multispeaker and Multilingual Scenarios.