IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 3, March 2026

DOI: 10.17148/IJARCCE.2026.153149

A Multimodal AI Framework for Real-Time Audience Engagement Detection in Virtual Communication

A V Tejaswi[1], M Sumana Sree[2], S Sahithi[3], Dr. C. Swapna[4]

Student, Dept. of AI & Data Science, Stanley College of Engineering and Technology for Women, Hyderabad, India[1-3]

Associate Professor, Dept. of AI & Data Science,

Stanley College of Engineering and Technology for Women, Hyderabad, India[4]

Abstract: The way we globally collaborate has been significantly impacted by virtual communication tools. However, these tools are not effective in conveying nonverbal cues of engagement, which are important in the process of effective human interaction. Real-time audience attention evaluation is a problem that faces the presenter in virtual meetings and classrooms. This study discusses the AI-based methodology that can be used in detecting audience engagement via multimodal emotion recognition.The interactive Speaker Dashboard also displays individual participant engagement scores, which are determined by a Node.js backend server running a Video Analyzer, an Audio Analyzer, and an Engagement Engine. This proposed framework is applicable in various scenarios, such as remote training sessions, corporate meetings, virtual classroom scenarios, and customer support interactions. This framework also indicates better detection accuracy. This sensitivity to real participant behaviour can be seen in the experimental results, which demonstrate that the level of engagement aligns with the attentiveness pattern of live sessions.

Keywords: Engagement detection; Multimodal emotion recognition; Facial expression recognition; Speech emotion analysis; Affective computing; WebRTC; Virtual collaboration; Deep learning; Attention tracking.

I. INTRODUCTION

The most revolutionary technology in the twenty-first century is artificial intelligence (AI). AI enables robots to perform activities that were previously considered to be the sole domain of the human mind, such as perception, reasoning, and emotions. The widespread adoption of AI-based tools in the realms of education, healthcare, business, and communication has created new avenues to deal with the problems that plague human interactions. One of the most significant is the maintenance of interaction in the virtual world.

The COVID-19 pandemic has accelerated the shift to remote working and online education across the world, and therefore, tools such as Zoom, Google Meet, and Microsoft Teams have become indispensable in day-to-day business and educational activities. However, it is imperative to note that the intricate network of non-verbal cues is an indispensable aspect of face-to-face interaction, which these digital devices are devoid of, despite their convenience and global accessibility. A speaker can always assess whether his/her audience is attentive, confused, bored, or disinterested through facial expressions, gestures, body positions, eye contact, and tone of voice.

Such cues are generally absent for the speaker in an online environment. Moreover, the speaker misses instant feedback, which would normally be available in a real environment due to factors such as camera positions, poor resolution of the video feed, or cameras not functioning at all. Thus, disengagement of the audience is generally not recognized in time, except when it significantly impacts the effectiveness of the training, meeting, or learning process [3].

The area of affective computing, which deals with computers that can recognize, interpret, and reproduce human emotions, has shown significant potential in reducing the gap. The way to implement an emotionally intelligent environment for virtual communications is through multimodal emotion recognition systems, which can simultaneously investigate speech, facial, and behavioral cues in real-time [4]. Despite this potential, most solutions available today focus on only one modality, like only facial expressions, which reduces their overall effectiveness and accuracy in real-world situations in which images may be partially obstructed, illumination may be insufficient, and audio equipment may not be functioning properly [5].

An AI-based framework for engagement detection in virtual meeting environments is proposed in this paper. The system utilizes a multimodal approach consisting of behavioral attention tracking, audio-driven speech emotion detection, and

DOI: 10.17148/IJARCCE.2026.153149

facial emotion recognition. This enables the system to provide dynamic and real-time engagement metrics that are directly displayed within the browser interface. The main contributions of this study are two-fold: (1) the development of a real-time multimodal engagement scoring pipeline, which utilizes both audio and visual modalities, and (2) the design of a reliability-weighted late fusion mechanism, which can function well even in scenarios where individual modalities are absent.

II. LITERATURE SURVEY

A. *Multimodal Emotion Recognition*

Using CNN, RNN/LSTM, and NLP with attention fusion in text, audio, and video modalities, Kannappan (2025) proposed emotion-aware AI systems for remote teams. This method reported 80% accuracy in text, 82% in speech, and 87% in facial accuracy[6]. This study proved that multimodal fusion always performs better than unimodal fusion. However, it also exposed some of its disadvantages, e.g., the need for enhancing data privacy guarantees and cultural adaptation of emotion models. Moreover, there is also the work of Singh (2025), which proposed VisioPhysioENet, a new architecture that integrates physiological signals collected using rPPG with visual behavioral cues such as gaze direction and body posture with 63.09% accuracy on the DAiSEE benchmark dataset[7]. This is especially noteworthy since it utilizes both visual and non-invasive physiological feedback. Additionally, Villegas-Ch (2025) developed a multimodal emotional detection system that is production-ready and works directly with Microsoft Teams via the Microsoft Graph API. In practical meeting scenarios, our system achieves precision values of up to 0.95 for positive and neutral emotion categories using both an LSTM for voice analysis and a CNN for facial features extraction. [8]

B. *Facial Expression-Based Engagement Detection*

By utilizing the ResNet-50 model that the authors have trained with the FER2013, RAF-DB, and CK+ datasets to detect learners' engagements, Gupta (2023) reported an impressive accuracy of 92.3%. The researchers proposed the new Engagement Index measure based on the probabilities of facial expressions in this study [9].

Further, the researcher Watanabe proposed a method for measuring learners' involvement in online meetings by leveraging the MobileNetV2 deep features and the OpenFace facial action units. The proposed approach attained an impressive F1-score of 0.895 via the leave-one-participant-out cross-validation approach [10].

In the same context, Irfan et al. proposed EngageSense, which utilizes CNN-based gaze detection, Dlib-based HOG+SVM-based face detection, and OpenPose-based body posture detection to measure learners' participation in online learning scenarios. Although the authors recognized the issues associated with the limited dataset, sensitivity to lighting, and the inability to sense physiological signals, this system was able to achieve an accuracy of 79.5% in detecting gaze [11].

C. *Audio and Conversational Analytics*

Vocal softness and whispering were found to greatly impact the perceived level of engagement, speech intelligibility, and pleasantness by Cordourier Maruri et al. in their methodical research on the impact of voice changes in the context of simulated virtual meetings with amplitude and pitch analysis [12]. Furthermore, Zaheer developed a critical theoretical basis for the development of AI-based emotional intelligence in virtual collaboration spaces with the evaluation of data collected from 200 remote workers and the establishment of strong statistically significant relationships with emotional intelligence and the efficacy of collaboration ($r = 0.57$) and the quality of team communication ($r = 0.62$) [13].

D. *Research Gaps*

Three important research deficits are identified through the synthesis of the reviewed literature. First and foremost, the majority of the existing algorithms are unimodal, using either voice recognition or facial expression recognition alone. This is seen to affect the robustness of the detection mechanism in real-life scenarios where one of the modalities is not available. Perhaps more importantly, however, is the fact that few algorithms were tested in real-life, live virtual meeting scenarios, with the majority being tested using controlled datasets in the lab. This is where the proposed framework excels.

III. PROPOSED SYSTEM

A. *System Overview*

A proposed architecture for a system that integrates WebRTC-based video conferencing and an AI-based backend for identifying multimodal emotion will be a comprehensive browser-based solution for engagement analysis. The system will analyze facial expressions and vocal patterns of the participants in real-time during the session, and engagement scores will be computed and displayed in real-time within the browser-based interface. No raw media will be stored,

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 3, March 2026

DOI: 10.17148/IJARCCE.2026.153149

ensuring that all participants are kept private. The participants can use any web browser to access the session, and no additional software installation is required.

B. *System Architecture*

The architecture, depicted in Fig. 1, organizes five closely integrated and independently maintainable layers. The Web Client, developed using React, functions as the primary user interface — it captures the webcam video and microphone audio of each participant and streams this data in real-time to the backend through secure WebSocket connections, while concurrently rendering live visualizations of engagement scores within the browser. The Backend Server, constructed in Node.js with Express, serves as the central orchestration hub: it receives incoming media streams, directs video data to the Video Analyzer and audio data to the Audio Analyzer, aggregates their respective outputs, and transmits the combined results to the Engagement Engine. The Video Analyzer employs face detection and facial landmark extraction, followed by CNN-based emotion classification, to yield a continuous video engagement score V(t). Moreover, the Audio Analyzer extracts acoustic features and utilizes an LSTM+SVM pipeline to generate a continuous audio engagement score A(t). The Engagement Engine integrates both scores using a reliability-weighted formula and produces the final normalized engagement metric.

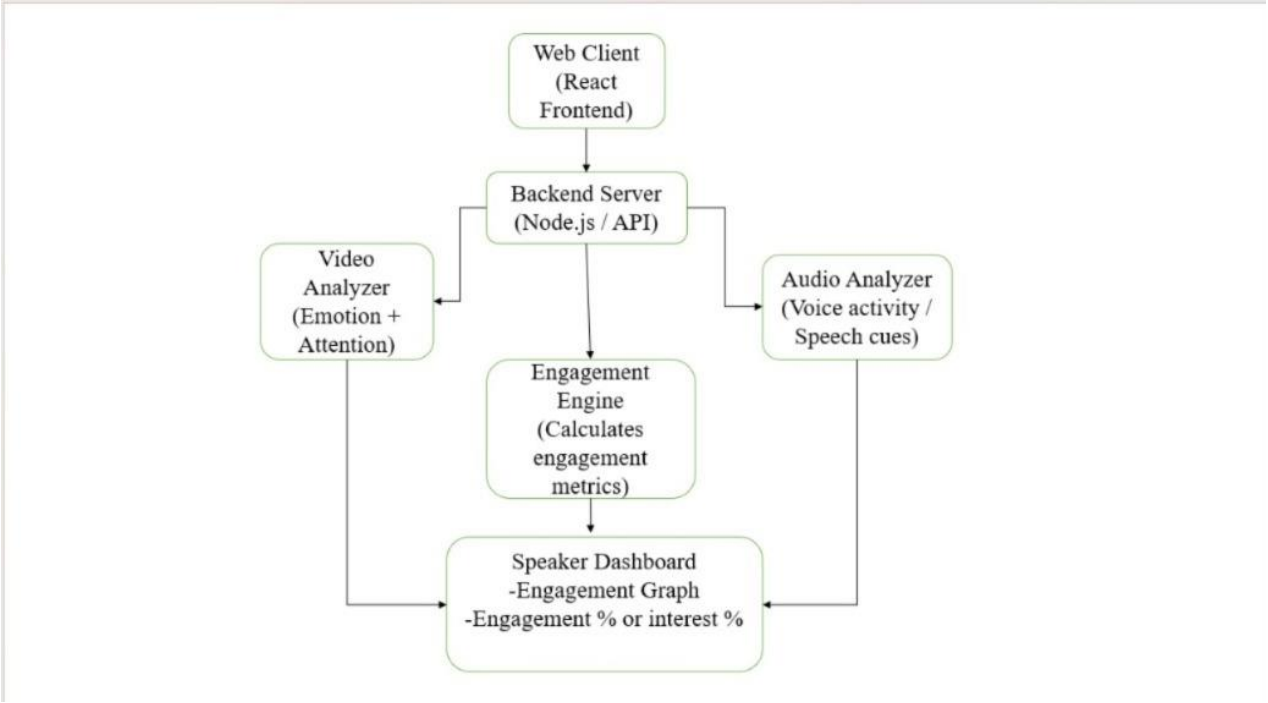


Fig. 1. Proposed Framework System Architecture

TABLE II SYSTEM ARCHITECTURE COMPONENTS

Component	Technology	Function	Output
Web Client	React, WebRTC, WebSocket	Media capture & real-time engagement display	Live video/audio streams
Backend Server	Node.js, Express	Stream orchestration & routing	Aggregated outputs
Video Analyzer	OpenCV, CNN, TensorFlow	Facial emotion & attention detection	Score V(t)
Audio Analyzer	Librosa, PyAudio, LSTM+SVM	Speech emotion classification	Score A(t)
Engagement Engine	Weighted late fusion	Score fusion & engagement output	E(t) = aV(t) + bA(t)

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 3, March 2026

DOI: 10.17148/IJARCCE.2026.153149

IV. METHODOLOGY

A. *Multimodal Data Acquisition*

The system uses two main sources of sensor data. First, it uses a video analysis pipeline to analyze webcam data frame by frame. It captures webcam data and plays it back at a customizable frame rate. In order to ensure temporal continuity in emotion tracking, it continuously buffers microphone data and divides it into overlapping segments of varying length. A key design choice is to ensure that it does not store audio buffers or video frames. All analysis is done in memory during the current session, and data is deleted at the end of the session. This is done in order to comply with data privacy laws. End-to-end encrypted peer-to-peer communication protected by protocols is facilitated by WebRTC. WebRTC is used to provide the real-time transport layer.

B. *Video-Based Emotion Detection*

The video recognition pipeline involves face detection through the Dlib HOG+SVM detector for accuracy and Haar Cascade for speed. The pipeline employs a facial landmark extractor that recognizes face areas identified by the face detection method. The extractor recognizes 68 key points on the face, including lips, nose, eyes, eyebrows, and jawline. These points are used by the system to recognize geometric features that describe face muscle configurations. The identified face is then classified into one of the five emotion states associated with face engagement through a Convolutional Neural Network (CNN) trained on benchmark datasets like FER2013 and AffectNet. The emotion states include attentive, neutral, perplexed, bored, and disengaged. To ascertain whether the participant is looking at the screen or turning away, the head posture estimation simultaneously calculates the three Euler angles, such as yaw, pitch, and roll. Moreover, the gaze direction estimate signal is also provided in the form of an additional attention signal. The emotion categorization confidence score and the attention signal derived from the head posture and gaze estimation are weighted and combined by the algorithm to arrive at the video engagement score V(t).

C. *Audio-Based Emotion Analysis*

For every segment of the speech that has been buffered, the process of feature extraction is comprehensive in the audio analysis pipeline. The spectral envelope of the speech signal is closely linked with the identity of the phoneme and the emotional state. The fundamental frequency, or the pitch F0, monitors the stress and the prosody. The volume and the level of engagement in the vocals are measured in terms of the Root Mean Square Energy (RMS). The features also include the spectral centroid, the pause ratio, and the speech pace. However, the Support Vector Machine classifier is used for the final classification of the emotion category. The LSTM network is used to analyze the features over time and detect the changes in the prosody and the emotion. The vocal energy, the emotional expressiveness, and the speech continuity are all combined into one normalized value in the audio engagement score A(t).

D. *Engagement Score Fusion*

Using the formula for reliability-weighted fusion, where E(t) is given by E(t) = a*V(t) + b*A(t), with non-negative weights that sum up to 1, the Engagement Engine uses late fusion of video and audio engagement scores. Although fusion with fixed weights is computationally simple, the system adjusts the weights according to signal quality and availability. The system sets weight 'a' for visual modality to 0 and weight 'b' for audio modality to 1 if the participant deactivates their camera, effectively allowing the system to run on audio alone. Conversely, if the microphone is muted, the system adjusts the weights in the opposite manner. Rather than causing complete system failure, the system now guarantees gradual deterioration. In order to remove the effects of the spikes and generate stable and comprehensible curves of the trends in the engagement levels normalized over the range 0-100%, the system uses a sliding window smoother on the fused signal E(t).

V. RESULTS AND DISCUSSION

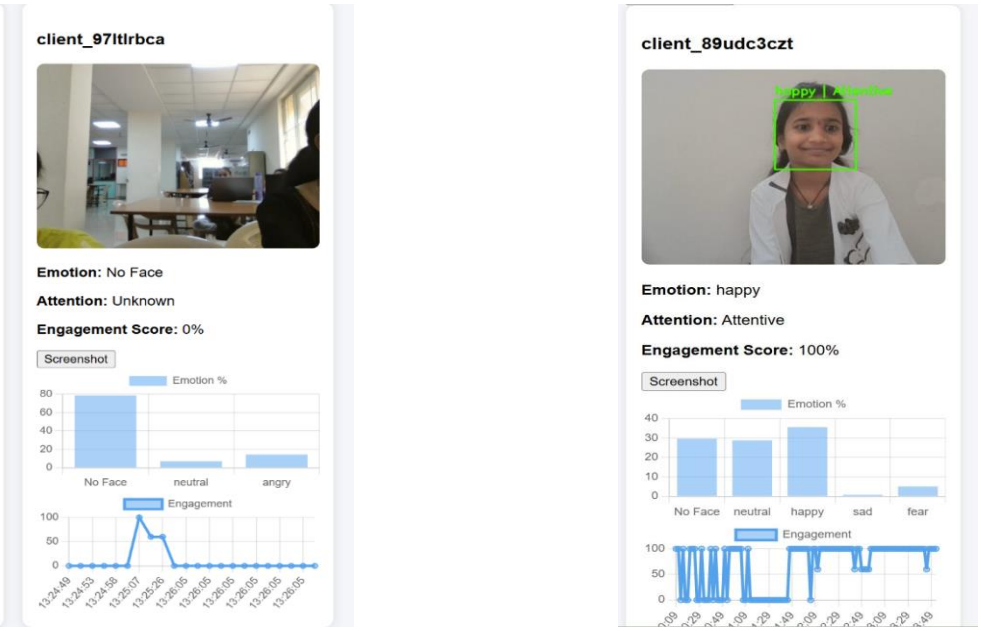
The proposed multimodal framework was implemented and tested in a virtual session in real time. The system's output for the participant "client_97ltlrbca" in a testing session, as shown in Figure 2, displays "No Face" if there's no face detected in the video frame, and "Unknown" if the gaze state cannot be determined. The "Emotion % Bar Chart" and "Engagement Timeline Graph" are two types of analytical visualizations provided by the proposed system. Another successful face detection scenario for the participant "client_89udc3czt" is shown in Figure 2, where the system successfully detects the participant's face in the video frame and accurately determines the participant's emotional state and engagement level.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 3, March 2026

DOI: 10.17148/IJARCCE.2026.153149



(a) client_97ltlrbca — No Face / Unknown (b) client_89udc3czt — Happy / Attentive / 100%

Fig. 2. Real-Time Framework Output Comparison — (a) client_97ltlrbca: No Face Detected, Unknown Attention; (b) client_89udc3czt: Happy Emotion, Attentive State, 100% Engagement Score

As for the distribution of recognized emotional states, it is presented in the Emotion % bar chart. This chart summarizes emotion states recognized in the session. The state "No Face" is dominant at around 75% in the first session (client_97ltlrbca), which means that at these times, the participant was out of the camera's frame. The state of being neutral is at around 8%, and furious is at around 15%.Moreover, this chart presents an important behavior-related result. The system does not misclassify times of physical disengagement as states of being neutral; instead, it recognizes them through the absence of face detection. The Emotion % bar chart for the participant "client_89udc3czt" in Figure 3, on the other hand, indicates that the framework was able to successfully detect the participant's face during the session and that the most dominant emotion was classified as "happy," making up the largest portion of the total detected emotional state at around 35%, followed by neutral and "No Face" at around 30%, and then fear at around 5%. In the event that the participant's attention state is classified as "Attentive," the system provides the participant with a real-time Engagement Score of 100%, reflecting that the participant is fully engaged. The participant in the scenario above was seen to have a high level of engagement over the entire recorded session, with the score consistently reaching 100% at regular intervals in the participant's engagement timeline graph. This outcome further verifies the effectiveness of the system in properly distinguishing between high levels of engagement and disengagement using the proposed framework without the need for manual annotations, which ensures the system's capacity to properly measure true student engagement when the participant is actively facing the camera with the right emotional disposition.

The computed engagement score E(t) is displayed on the graph of the engagement timeline between the times 13:24:49 and 13:26:05. As seen on the graph, the levels of involvement are near zero most of the time, but there is a notable spike at 100% at time 13:25:07, which immediately returns to near zero levels at around 13:25:26 after a sharp drop from the peak of 100% involvement, and then returns to near zero levels again. This temporal pattern is closely related to the actions of the participant, where at around the time of the spike at 13:25:07, the subject was seen facing the camera and showing signs of high involvement before disengaging again. This shows the effectiveness of the pipeline's real-time accuracy, where the system is able to compute the involvement score at the time the subject was seen facing the camera.

The main drawback, as identified in previous studies, is effectively overcome by the proposed multimodal fusion technique. The video modality's score returns to zero if the participant leaves the frame. In this situation, the proposed dynamic weight adjustment technique adjusts the value of b, increasing it and correspondingly decreasing the value of a, so that the audio modality still contributes to E(t). This is in line with the accuracy improvements of 5-12% for multimodal systems over unimodal systems, as reported by Villegas-Ch (2025) and Kannappan (2025) [6][8]. The proposed framework thus ensures that engagement is always monitored, even in situations where individual modalities may not be working, which is a necessary condition for the successful deployment of this technique in the real world.

International Journal of Advanced Research in Computer and Communication Engineering
Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 3, March 2026
DOI: 10.17148/IJARCCE.2026.153149

Moreover, a comparison with previous works would show the advantages of the proposed paradigm. The proposed system can process unstructured live video streams with various illumination, camera positions, and natural participant behaviors, whereas Gupta (2023) and Watanabe (2023) can only attain high accuracy on structured, pre-recorded facial expressions dataset [9][10]. The proposed framework can use common webcam and microphone interfaces, which are available on any modern device, despite the contributions of Irfan et al. (2024), which requires a specific eye image dataset and completely neglects audio analysis [11]. Since none of the studies address the issue of graceful modality deterioration in live sessions, the proposed approach of dynamic reliability-weighted fusion by the Engagement Engine is a new contribution in this area.

The proposed approach is compared with six existing systems in Table III in terms of various evaluation factors such as modality, dataset, accuracy, real-time capability, and critical restrictions.

TABLE III COMPARISON WITH EXISTING ENGAGEMENT DETECTION SYSTEMS

System	Modality	Dataset	Accuracy	Real-Time	Limitation vs Proposed
Kannappan 2025	Video+Audio+Text	Custom multimodal	87% (face)	No	No dynamic weight adjustment; not deployed live
Gupta 2023	Video only	FER2013, RAF-DB	92.3%	Partial	Unimodal; fails when camera off
Watanabe 2023	Video only	24 participants	F1=0.895	Yes	No audio; controlled lab only
Irfan et al. 2024	Video only	4,453 eye images	79.5% gaze	Yes	Needs special dataset; no audio modality
Villegas-Ch 2025	Video+Audio	AffectNet,RAVDESS	Prec. 0.95	Yes	Platform-dependent (MS Teams API only)
Singh 2025	Video+Physio	DAiSEE	63.09%	No	Needs physiological sensors; low accuracy
Proposed Framework	**Video+Audio**	**Live sessions**	**Real-time**	**Yes**	**Dynamic weight fusion; browser-based; no special hardware**

The scalability and modularity of this framework have also been qualitatively evaluated. The fact that all components of the pipeline can be individually upgraded without affecting the rest of the pipeline follows from the fact that the Video Analyzer, Audio Analyzer, and Engagement Engine are separate modules that communicate with each other through the Node.js backend. For example, the developers can replace the CNN-based emotion classifier with a transformer-based classifier that was trained on a larger and more diverse dataset without making any modifications to the audio pipeline or the Engagement Engine. The modularity of the pipeline also facilitates extensions to additional modalities. For example, if the developers wanted to add a physiological signal module, they would only need to modify the Engagement Engine's weights and add a new analyzer component. Due to the independent, asynchronous processing of each participant's stream, the WebRTC communication layer guarantees that the framework can be expanded to accommodate a multi-participant group session.

The framework is limited by the computational capabilities of standard consumer-grade hardware. It can adjust the balance between detection granularity and CPU load, depending on the environment, because of the variable frame rate of the video processing pipeline. It is not necessary to store large volumes of data because of the short periods of the audio segmentation, which can capture sudden changes in emotion. The asynchronous event-driven design of the Node.js framework makes it suitable for real sessions with multiple, concurrent participants by preventing delays in the processing of one participant's stream from interfering with the concurrent processing of other streams.

International Journal of Advanced Research in Computer and Communication Engineering
Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 3, March 2026
DOI: 10.17148/IJARCCE.2026.153149

VI. CONCLUSION AND FUTURE SCOPE

In order to address the basic issue of engagement visibility in virtual communication, this study proposes an artificial intelligence-based paradigm for engagement detection that incorporates real-time face expression detection, speech emotion detection, and behavioral attention tracking.

The React Web Client, Node.js Backend Server, Video Analyzer, Audio Analyzer, and Engagement Engine form the modular five-layer architecture of the proposed system, which analyzes multimodal inputs from standard webcam and microphone inputs to derive a continuous, normalized engagement score for all participants. Moreover, the reliable weighted late fusion approach makes the proposed system unique in comparison with previous single-modality-based approaches in that it guarantees accurate engagement tracking during sessions even in cases of partial signal failure, where participants deactivate their cameras or microphones.

Also, the evaluation of the live system showed that the engagement timeline effectively mirrors the attentiveness behaviors of real participants, with the system skillfully detecting short engagement intervals and long disengagement phases, including the absence of the camera.The graphic of Emotion % helps to improve the real-time timeline for reflection by providing a summary of the session-level behaviors.

The following are the areas that need to be addressed in the future: (1) the emotion recognition module has to be enhanced to include the recognition of micro-expressions and levels of cognitive load and fatigue; (2) the inclusion of more behavioral modalities, such as eye-tracking and gesture recognition, has to be considered to draw more accurate inferences; (3) the development of extensive analytics dashboards has to be undertaken to support the analysis of user engagement trends throughout multiple sessions; (4) the development of native plugin versions of tools such as Zoom, Google Meet, and Learning Management Systems has to be undertaken to reduce the barriers to adoption; and (5) extensive user studies need to be undertaken using ground truth annotations gathered from professional and student populations to validate the accuracy of the system against diverse demographics, cultures, and scenarios.

The practical implications of this approach extend far beyond the use case of virtual meetings themselves. The reliability-weighted late fusion architecture can be used as a flexible design pattern in any multimodal signal processing system where individual sensors might be expected to fail. The potential future applications of this approach include telehealth consultations, public speaking coaching systems, virtual interview assessment systems, intelligent classrooms, and anywhere else where real-time emotional and attentional awareness significantly improves the quality of interaction. The design of the application is significantly easier to adopt compared to native application alternatives, as it is browser-based and doesn't require installation, which makes it easier to implement without requiring significant changes in IT infrastructure in corporate training environments, educational institutions, and customer support environments.

To summarize, the proposed AI-based method for the detection of involvement in the above text outlines a promising new way forward for the development of effective, flexible, and emotive virtual communications. The proposed methodology for the development of engagement-aware virtual systems improves the technological design and usability of such systems by considering the detection of engagement in a system as a process of continuous, reliability-weighted estimation, rather than a binary classification process. The proposed work provides a sound foundation for future researchers to build on to develop complex, flexible, and human-centric virtual communications by incorporating real-time scoring for all participants, modality weight modifications, and an extensible design. The increased need for efficient collaboration tools in various fields, including business, training, healthcare, and education, further indicates that it is imperative to provide these tools, thus making it not only desirable but also necessary.

REFERENCES

[1]. A. A. Burcea, "Emotional Economy in the Digital Age: How Conversational Analytics Shapes Team Morale in Collaborative Platforms," Ovidius University Annals, Economic Sciences Series, vol. XXV(1), 2025.
[2]. S. Zaheer, "The Role of Emotional Intelligence in Remote Team Collaboration and Communication," Journal of Personnel Management, 2024.
[3]. M. A. A. Dewan, M. Murshed, and F. Lin, "Engagement Detection in Online Learning: A Review," Smart Learning Environments, 2019.
[4]. B. A. Erol, A. Majumdar, and P. Benavidez, "Toward Artificial Emotional Intelligence for Cooperative Social Human-Machine Interaction," IEEE Trans. Comput. Social Syst., vol. 7, 2020.
[5]. C. Audrin and B. Audrin, "More Than Just Emotional Intelligence Online: Introducing Digital Emotional Intelligence," Frontiers in Psychology, 2023.

International Journal of Advanced Research in Computer and Communication Engineering
Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 3, March 2026
DOI: 10.17148/IJARCCE.2026.153149

[6]. S. Kannappan, "Emotion Aware Artificial Intelligence Systems for Future Remote Team," ICSICE 24, vol. 120, 2025.
[7]. A. Singh, "VisioPhysioENet: Visual Physiological Engagement Detection Network," Elsevier Preprint, Aug. 2025.
[8]. W. Villegas-Ch, "Multimodal Emotional Detection System for Virtual Educational Environments: Integration into MS Teams," IEEE Access, 2025.
[9]. S. Gupta, "Facial Emotion Recognition for Real Time Learner Engagement," Multimedia Tools and Applications, vol. 82, 2023.
[10]. K. Watanabe, "EnGauge: Engagement Gauge of Meeting Participants Estimated by Facial Expression and Deep Neural Network," IEEE Access, 2023.
[11]. M. Irfan, P. Patel, and B. Hassan, "Engage Sense: A Hybrid Approach for Real Time Engagement Detection for Virtual Classrooms," IEEE EDUCON, 2024.
[12]. H. A. Cordourier Maruri, S. Aslan, and G. Stemmeri, "Analysis of Contextual Voice Changes in Remote Meetings," INTERSPEECH, 2021.
[13]. A. Levordashka, D. S. Fraser, and I. D. Gilchrist, "Measuring Real-Time Cognitive Engagement in Remote Audiences," Scientific Reports, 2023.
[14]. M. Rahman, M. Ahmed, and M. Mahmud, "Assessing Participants' Engagement in Virtual Meetings Using Facial Landmarks and Deep Learning," ICCA, 2024.
[15]. S. Mandia et al., "EngageFormer: Transformer-Driven Modeling for Classifying Student Engagement in Online Learning," 2023.

