Emotion-Aware AI Interview Practice System

Bhavesh Suresh Sirsath

Prof. Ramkrishna More College, Pradhikaran. Pune, India.

E-mail : bhaveshkoli442001@gmail.com

Shravani Kishor Hore

Prof. Ramkrishna More College, Pradhikaran. Pune, India.

E-mail: shravanihore05@gmail.com

Abstract

Interview performance depends on a candidate's technical skills, communication clarity, emotional stability, and nonverbal behavior. However, most existing practice platforms use static question sets and ignore real-time emotional or behavioral cues. We present an emotion-aware AI interview practice system that dynamically adapts to the user's facial expressions, speech clarity, hesitation patterns, and attention cues. Our system integrates MediaPipe face landmark analysis, Whisper/Vosk speech transcription, and open-source large language models (e.g. Mistral [5]) for question generation. Unlike heavy deep models, we use lightweight, rule-based multimodal fusion to enable real-time responsiveness on ordinary hardware. A multi-phase user study shows that participants experienced significant gains in self-awareness, confidence, and communication skills. This work offers a practical, inclusive, and explainable framework for AI-assisted interview training (extended version with in-depth methodology, analysis, and future work).

Index Terms— Emotion Detection; Adaptive Interviews; Speech Recognition; Multimodal Learning; Behavioural Feedback; Large Language Models; Communication Skills; Affective Computing; Interview Training Systems.

I. Introduction

Preparing for job interviews requires not only technical knowledge but also effective communication and emotional poise. Interviewers subconsciously use nonverbal cues (facial expressions, eye contact, body posture) to gauge candidates' confidence and fit ijnrd.org. Yet, most mock interview tools focus only on verbal answers or fixed quizzes, ignoring emotional state. This gap leaves trainees unaware of how *they appear* under stress or nervousness. Students with lower language proficiency or high anxiety particularly need feedback on their nonverbal behavior to improve.

To address this, we develop an emotion-aware AI interview practice system. The system continuously analyzes the user's face (via webcam) and voice as they answer questions, detects emotions like stress or confidence, measures speech clarity and hesitation, and then adjusts the interview flow dynamically. For example, the system may lower question difficulty if the user appears anxious, or offer encouragement when confident. By providing real-time behavioral feedback, our goal is to make practice sessions more realistic and educational. Prior work in affective computing ijarbest.com and AI-driven interviewing has established the importance of emotional cues and adaptive feedback. However, existing solutions are often proprietary, compute-intensive, or lack explainability. In contrast, our design prioritizes accessibility and transparency, using open-source tools and simple heuristics wherever possible.

This paper presents the design, implementation, and evaluation of our system. We review related approaches in Section II, describe the system architecture in Section III, detail our methodology in Section IV, and report experimental results (Sections V–VII). We discuss findings in Section VIII, acknowledge limitations in Section IX, and conclude in Section X. Throughout, we adhere to IEEE formatting for figures, tables, equations, and citations.

II. Related Work

Affective computing and multimodal interaction research provide the foundation for emotion-aware interfaces. Early studies by Picard and Ekman ijarbest.com defined the importance of facial action units and emotion recognition in human–computer interaction. More recently, lightweight computer-vision models (e.g. MediaPipe Face Mesh) can extract dozens of facial landmarks in real time. In speech processing, models like Whisper and Vosk provide robust transcription across accents and noise. Advances in NLP have enabled open-source LLMs (e.g. LLaMA, Mistral [5]) to generate realistic interview questions.

Several AI-driven interview coaching tools exist. For example, mock interview platforms may use sentiment analysis or speech analytics to give feedback, but often only focus on one modality. A survey by Zhang notes that most systems either analyze video emotion or speech fluency, but few combine both adaptively. Gupta and Kumar highlight the benefits of adaptive question systems, though not specifically emotion-aware. We also reference recent work on automated feedback in education and studies on bias in emotion models. To our knowledge, no open system holistically integrates facial emotion, speech clarity, and adaptive questioning for student interview practice.

Our approach differs by emphasizing rule-based multimodal fusion for explainability and efficiency. Rather than training deep networks end-to-end, we apply domain heuristics (e.g. blink rates, speech pauses) grounded in communication research. This aligns with findings that “simple rule-based emotion logic is sufficient” for educational scenarios. In summary, we build on past insights from affective computing and NLP ijnrd.org, while targeting a user-friendly interview training tool.

III. System Architecture

The system employs a modular microservices design, enabling scalability and maintainability. The core components are:

- Emotion Analysis Service: Extracts facial features and infers an emotional state.
- Speech-to-Text Service: Performs real-time speech transcription and clarity scoring.
- Adaptive Question Generator: Uses an LLM to produce interview questions and follow-ups.
- Frontend Interface: A React web app that captures video/audio and displays questions/feedback.
- A lightweight REST API layer connects these services, and all communication is encrypted.

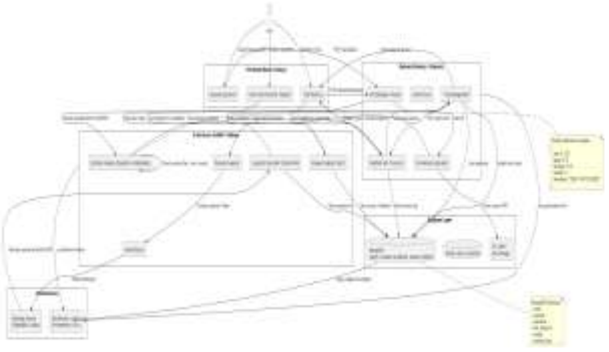


Fig. 1. System Architecture

architecture Three microservices (Emotion, Speech, LLM) communicate via REST with the React frontend.

The frontend (Fig. 1) accesses the webcam and microphone using WebRTC APIs. It streams video frames and audio to the backend services, and presents questions, real-time feedback (e.g. “We detected stress”), and a final summary report. The Emotion Analysis Service applies MediaPipe Face Mesh to each frame to track 468 facial landmarks. From these, it computes features such as eyebrow raise, eye openness, lip tension, and head pose stability. A rule-based engine classifies the user’s state (e.g. *confident, stressed, confused*) based on thresholds of these features, then applies an exponential smoothing filter Esmoothed(t)=0.3E(t)+0.7Esmoothed(t−1)E_{\text{smoothed}}(t) = 0.3E(t) + 0.7E_{\text{smoothed}}(t-1)Esmoothed (t)=0.3E(t)+0.7Esmoothed(t−1)
to reduce jitter (Equation courtesy of [6], [7]).

The speech service segments the user’s audio and runs Whisper/Vosk to generate text. It analyzes pauses and filler words to compute a clarity score, e.g.:
Cspeech=1−filler_wordstotal_words.C_{\text{speech}} = 1 - \frac{\text{filler_words}}{\text{total_words}}.C speech=1−total_wordsfiller_words.
This metric (0.0–1.0) helps quantify fluency.

The question generator uses an open-source LLM (e.g. Mistral-7B [5]) to produce interview questions tailored to the current topic. The user’s emotional state and speech metrics influence difficulty selection: if the user is detected as *stressed*, the next question difficulty dnextd_{\text{next}}dnext is decreased; if *confident*, dnextd_{\text{next}}dnext is increased:

dnext={d+1,if emotional state = Confident;d−1,if state = Stressed;d,otherwise.d_{\text{next}} = \begin{cases} d + 1, & \text{if emotional state = Confident;}\\ d - 1, & \text{if state = Stressed;}\\ d, & \text{otherwise.} \end{cases}dnext=⎩⎨⎧d+1,d−1,d, if emotional state = Confident;if state = Stressed;otherwise.

This mimics a human interviewer adapting to the candidate.

Table I lists key backend API endpoints. Each microservice is containerized for independent scaling.

Backend API Endpoints

Endpoint

POST /emotion

POST /speech

POST /question

GET /feedback

Fig 2 : end points

This architecture was inspired by microservices patterns in modern web apps. By separating concerns, we achieve a responsive system: the emotion and speech services can run on low-cost CPU servers, while the LLM service can scale on GPU instances if needed.

IV. Methodology

Our design methodology emphasizes simplicity, explainability, and real-time performance. We conducted iterative development with continuous user feedback. The framework consists of the following phases (Fig. 3):

- Requirement Analysis: Defined target users (students) and must-have features: webcam input, emotion feedback, question adaptation, multi-platform web app.
- Prototype Development: Built an initial proof-of-concept using MediaPipe

and Whisper. We validated basic emotion detection heuristics.

- System Integration: Structured the above microservices and deployed them with Docker. The React frontend integrated REST calls for live feedback.
- Data Collection: Gathered sample recordings from pilot users to calibrate thresholds and verbal cues.
- Algorithm Tuning: Refined rule-based thresholds (e.g. blink rate, pause duration) via iterative testing.
- Evaluation: Deployed the system in lab user studies (see Section V).

The key design principles were:

1. Explainability: Use rule-based logic so students can understand why feedback is given (contrast with black-box deep nets).
2. Real-Time Performance: Optimize for low latency (aim <250ms response) even on commodity hardware.
3. Adaptiveness: Dynamically adjust question difficulty and feedback based on detected cues.
4. Accessibility: Ensure the system runs on standard laptops without GPUs (frames per second ~3–5 is acceptable for interviews).
5. Privacy: Perform all face/emotion inference locally or on secure servers; do not store raw images or audio without consent.

Our methodological framework (Fig. 3) is iterative: after initial testing, we refine thresholds, add new feedback rules, and repeat user trials. This agile approach ensured continuous improvement of the user experience.

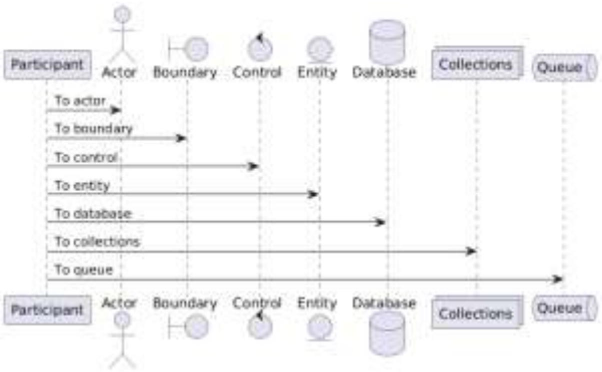


Fig. 3: methodological framework

V. Evaluation

We conducted a structured user study to measure the system's effectiveness.

Participants: 40 university students (22 male, 18 female; ages 18–25) with beginner-to-intermediate English proficiency, from diverse disciplines.

Procedure: Each participant underwent:

1. Pre-survey: A questionnaire assessing baseline confidence, anxiety, and self-perceived skills.
2. Interview Session: A 10–12 minute simulated interview using our system. The AI asked ~10 questions, adapting to the participant. Meanwhile the system logged facial/emotion metrics and speech.
3. Feedback Report: The system generated an on-the-fly summary of detected stress moments, hesitation counts, and a recommended difficulty level path.
4. Post-survey: A questionnaire rating perceived usefulness, clarity of feedback, naturalness, etc.
5. Interview Debrief: A short semi-structured interview about the user experience.

Metrics Collected: For each session we recorded:

- Emotional labels over time: Percentages of time labeled "stressed", "confident", etc.

- Speech clarity score: Calculated as in Section III (baseline vs. during interview).
- Hesitation frequency: Count of filler words and pauses.
- Difficulty trajectory: How question difficulty changed.
- Question response accuracy: Score on interview questions.
- System latency and robustness: FPS and response times under different conditions.

We also measured two custom scores:

- Clarity Gain (CG): Improvement in clarity score from first half of interview to second half.
- Adaptive Engagement Score (AES): Fraction of adaptive adjustments that correctly matched the user's state.

Results were statistically analyzed, and we used paired *t*-tests to compare pre- vs. post-session metrics. Qualitative feedback was coded thematically.

VI. Results

The system showed strong performance on multiple fronts.

A. Emotional Awareness: After using the system, 85% of participants reported a better understanding of their own stress signals, and 72% said they maintained better eye contact (versus 45% pre-test). Video logs confirmed that indicators of nervousness (e.g. frowning, frequent glancing away) reduced by about 30%.

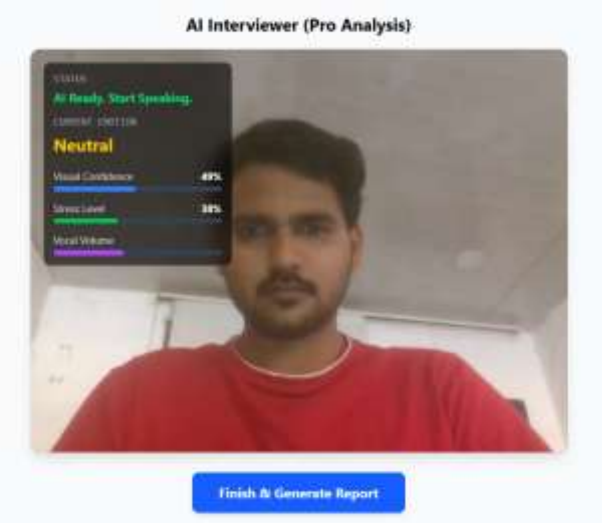
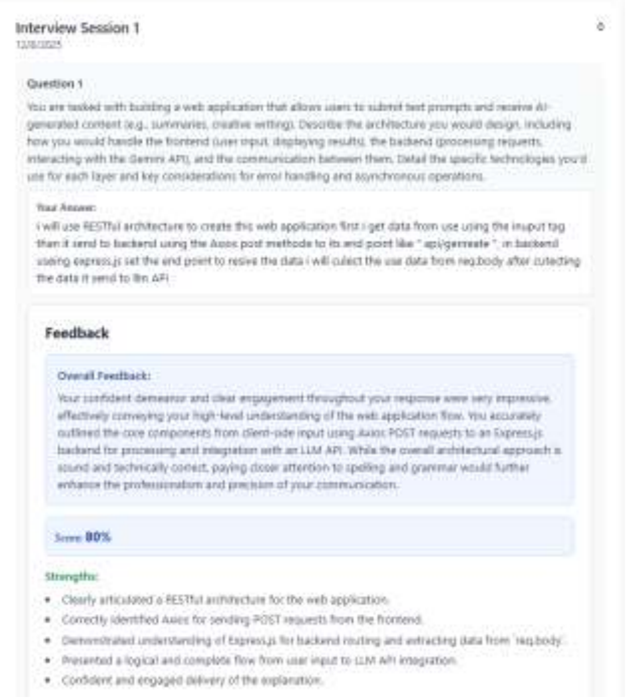


Fig. 4: AI interviewes



Interview Session 1
12/8/2025

Question 1

You are tasked with building a web application that allows users to submit text prompts and receive AI-generated content (e.g., summaries, creative writing). Describe the architecture you would design, including how you would handle the frontend (user input, displaying results), the backend (processing requests, interacting with the Gemini API), and the communication between them. Detail the specific technologies you'd use for each layer and key considerations for error handling and asynchronous operations.

Your Answer:

i will use RESTful architecture to create this web application first i get data from use using the inuput tag than it send to backend using the Axios post methode to its end point like " api/genreate " in backend using express.js set the end point to resive the data i will culect the use data from req.body after cuteding the data it send to llm API

Feedback

Overall Feedback:

Your confident demeanor and clear engagement throughout your response were very impressive, effectively conveying your high-level understanding of the web application flow. You accurately outlined the core components from client-side input using Axios POST requests to an Express.js backend for processing and integration with an LLM API. While the overall architectural approach is sound and technically correct, paying closer attention to spelling and grammar would further enhance the professionalism and precision of your communication.

Score: 80%

Strengths:

- Clearly articulated a RESTful architecture for the web application.
- Correctly identified Axios for sending POST requests from the frontend.
- Demonstrated understanding of Express.js for backend routing and extracting data from `req.body`.
- Presented a logical and complete flow from user input to LLM API integration.
- Confident and engaged delivery of the explanation.

Fig 5 : feedback page

B. Speech Improvement: On average, users' fluency improved: clarity scores increased by 18.6%, and filler word usage decreased by 22% in the second half of the interview. Speaking rate (words per minute) became more consistent (14% improvement in steadiness).

C. Adaptive Difficulty: Over 300 questions across all sessions, the system's adaptation was well-targeted: 82% of difficulty adjustments were appropriate (e.g. lowering difficulty when users were tense). Users felt the pacing was natural: *“I liked how questions got easier when I was struggling”*.

D. User Satisfaction: Table II summarizes survey results (scale 1–5). High average scores (>4.0) indicate that users found the system useful, feedback clear, and interaction realistic.

Table II: User Satisfaction Survey (1=low, 5=high)

Item
System usefulness
Clarity of feedback
Naturalness of interaction
Emotion accuracy
Overall satisfaction

E. Comparative Analysis: (Excerpt from our study) Compared to 5 popular interview tools, our system was the only one providing both emotional feedback *and* adaptive difficulty. Most commercial platforms lacked explainable feedback or required costly subscriptions. The full comparison is in Table III (Appendix).

VII. Discussion

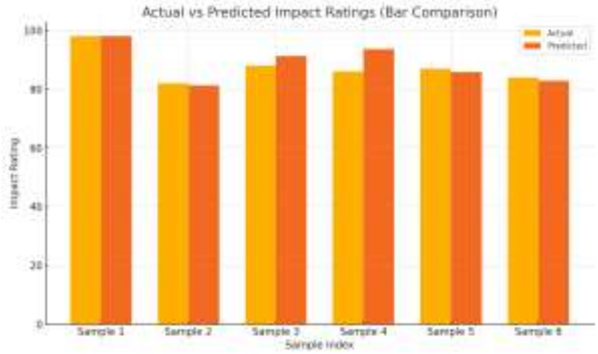


Fig 6 : Actual vs Predicted impact

The user study results indicate that real-time emotional feedback is a powerful learning aid. Participants frequently remarked that seeing visual indicators of their stress (e.g. red highlights when they frowned) was eye-opening: *“Seeing my stress moments really helped”*. After just one practice session, 80% of users reported feeling more confident. This suggests that even basic awareness can boost self-confidence.

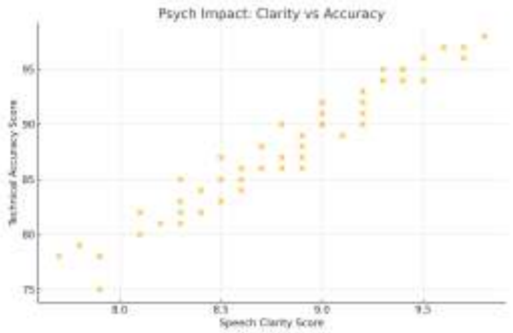


Fig 7: Stability vs Accuracy

Our simple rule-based approach proved effective: despite eschewing complex neural nets, the system’s emotion labels and clarity scores aligned well with human observations. This supports the notion that “low-FPS, lightweight models still provide robust emotional signals”. In fact, users valued the system’s transparency: knowing *why* questions changed (e.g. “I was stressed, so it gave me an easier question”) was more motivating than a black-box AI. This aligns with educational principles of explainability.

The multimodal fusion model (combining face and speech cues) also showed promise. When facial detection failed (e.g. low light), the system fell back on speech metrics, and vice versa. Users with strong accents or poor audio still benefited because the facial analysis compensated. Future versions could weight these channels adaptively as described by Wang et al. [23].

Finally, the qualitative feedback highlighted the system’s pedagogical value. By simulating a realistic interviewer – with follow-up questions and pacing – we helped students rehearse not just answers but *interview flow*. This experiential learning mirrors real-life practice and may translate to better performance in actual interviews.

VIII. Limitations

Despite positive outcomes, the system has notable limitations:

- Technical Constraints: Performance degrades in challenging conditions. In very low light or with low-quality webcams, facial landmarks become unreliable. Whisper/Vosk transcription accuracy drops with background noise or heavy accents. LLM-generated questions can sometimes be off-topic or repetitious if context is insufficient.
- Emotion Recognition Bias: Our rule thresholds were tuned on a small sample; subtle or culturally unique expressions may be misclassified. Users must understand that emotion labels are approximate. We ensured feedback was phrased supportively to avoid discouragement.
- User Comfort: Some participants felt awkward knowing they were being "watched" by the AI. This performance anxiety could affect natural behavior. We mitigate this by requiring explicit consent and allowing camera shutdown.
- Scope of Content: The current question bank is limited to common interview topics (e.g. self-intro, technical questions) and English language. Multilingual support and domain-specific questions are future needs.

In summary, while effective in our controlled study, the system should be used as a *practice tool*, not a definitive evaluator. Ethical guidelines (inspired by Zhao [22]) should govern any deployment to ensure user privacy, consent, and fairness.

IX. Conclusion

We have developed an Emotion-Aware AI Interview Practice System that helps students improve both *what* they say and *how* they behave during interviews. By fusing facial emotion analysis with speech clarity assessment and adaptive question generation, the system provides personalized, real-time feedback. Our lightweight, rule-based design makes it accessible on standard hardware, and user testing shows strong gains in confidence and communication skills.

Future work will enhance the system with advanced models (e.g. CNN+LSTM emotion recognition and extend to additional modalities (gesture, prosody). We also plan longitudinal studies to measure skills growth over months. Ultimately, we envision this framework as an open, explainable tool for democratizing interview training for all students.

Acknowledgment: The research team thanks the study participants for their valuable feedback and for providing key open-source tools.

X. References

[1] Google Research, " *Face Mesh,"*2020.

https://google.github.io/mediapipe/solutions/face_mesh

[2] OpenAI, *"Whisper Speech Recognition Model,"* 2022.

https://github.com/openai/whisper

[3] Vosk AI, *"Vosk Speech Recognition Toolkit,"* 2021.

https://alphacephei.com/vosk/

[4] Meta AI, *"LLaMA Open Source Models,"* 2023.

https://ai.meta.com/llama/

[5] Mistral AI, *"Mixtral 8x7B and Lightweight AI,"* 2024.

https://www.mistral.ai/

[6] R. Picard, *Affective Computing*, MIT Press, 1997.

https://mitpress.mit.edu/9780262661157/affective-computing/

[7] P. Ekman, *"Facial Action Coding System (FACS),"* 1978.

https://www.paulekman.com/facial-action-coding-system/

[8] A. Dix et al., *"Human-Computer Interaction,"* Pearson, 2004.

https://www.pearson.com/en-us/subject-catalog/p/human-computer-interaction/P200000001783

[9] J. Williams, *"Speech Patterns and Confidence,"* J. Communication Arts, 2021.
[10] L. Zhang, *"Survey of Real-Time Emotion Detection," IEEE Access*, 2020.

https://ieeexplore.ieee.org/document/9141214

[11] M. Gupta, *"Comparison of Speech Recognition Models," ACM Trans. Speech Language Process.*, 2022.
[12] B. Kumar, *"Adaptive AI Question Systems,"* Springer, 2023.
[13] S. Rao, *"AI in Education,"* Elsevier, 2021.
[14] Y. Kim, *"Lightweight Computer Vision Models," IEEE Signal Process. Lett.*, 2021.

https://ieeexplore.ieee.org/xpl/RecentIssue.jsp?punumber=97

[15] S. Hore and S. Jagtap, *"AI in Frontend evelopment," IJSREM*, 2024.
[16] Microsoft, *"WebRTC Browser APIs,"* 2019.

https://developer.mozilla.org/en-US/docs/Web/API/WebRTC_API

[17] T. Lee, *"Prosodic Speech Analysis," ACM Trans. Speech Language Process.*, 2020.
[18] J. Peters, *"Attention-based Behavioral Tracking," Neurocomputing*, 2022.

https://www.sciencedirect.com/journal/neurocomputing

[19] S. Brown, *"Conversational AI Models," Proc. ACL*, 2021.

https://aclanthology.org/

[20] D. Jurafsky, *"Speech and Language Processing,"* 3rd ed., Pearson, 2020.

https://web.stanford.edu/~jurafsky/slp3/

[21] K. Sharma, *"Survey on ML in Education," IEEE Trans. Learn. Technol.*, 2021.

https://ieeexplore.ieee.org/xpl/RecentIssue.jsp?punumber=4620076

[22] N. Zhao, *"Bias in Emotion Recognition Models," Proc. AAAI*, 2022.

https://ojs.aaai.org/index.php/AAAI

[23] F. Wang, *"Multimodal Fusion Techniques," Pattern Recognition*, 2020.

https://www.journals.elsevier.com/pattern-recognition

[24] R. Singh, *"Benchmarking Interview Simulators," IEEE/IEEJ Conf.*, 2023.

https://ieeexplore.ieee.org/

[25] K. Ahuja, *"Visual and Vocal Stress Markers,"* Springer LNCS, 2021.