Interview success with AI assistance

Surabhi Talekar[1], Sayali Landge[2], Prabha Gawde[3], Pruthvi Patil[4], Archana Salaskar[5]
[1,2,3,4]*Department. of Information Technology Vasantdada Patil Pratishthan's College of Engineering and Visual Arts*
[5]*Professor, Vasantdada Patil Pratishthan's College*

***Abstract*—Getting ready for interviews is something most students have to deal with at some point, especially when campus placements come around. While mock interviews and coaching sessions do help, not every student has access to them — costs can be high and finding the time is another challenge altogether. On top of that, the kind of feedback a student receives tends to vary from one interviewer to another, making it hard to track whether any real improvement is happening.**

We built a fairly straightforward AI-based mock interview system that lets students rehearse on their own, without needing anyone else in the room. The system records both video and audio during a session. It looks at the candidate's facial expressions through the video feed, and separately converts what they say into text to check how relevant their answers are. All of this gets pulled together into a single performance score at the end.

We weren't trying to replicate a real interview panel — the goal was simply to give students a reliable way to practice and get a sense of where they stand. Even with a relatively basic setup, the system turned out to be useful for giving candidates a clearer picture of their overall performance.

***Index Terms*—mock interview, sentiment analysis, facial expression recognition, performance score, artificial intelligence.**

I. INTRODUCTION

Interviews are a big deal — they are usually the final hurdle in getting a job or internship. A lot of students know their subject well but still freeze up or underperform when it actually matters, mostly because they haven't had enough chances to sit through something that feels like a real interview

Mock interviews are the go-to solution, but getting them arranged consistently is easier said than done. Many students end up leaning on friends or faculty, which is helpful but not always ideal — the feedback can be uneven, and it's hard to know if you're actually improving.

Thanks to recent progress in AI, there are now tools that can step in here. Facial expression analysis and speech processing, for instance, can give a fairly reasonable read on how someone is performing — not perfectly, but well enough to be genuinely useful.

In this project, an attempt is made to develop a simple web-based mock interview assistant. The system evaluates both how a candidate answers a question and how they present themselves while answering. The idea is to give users a rough but useful understanding of their performance so that they can improve with practice.

II. RELATED WORK

Facial expression recognition (FER) and automated affect analysis have been widely studied in the field of computer vision. Early work by Pantic and Rothkrantz [4] provided a detailed survey of automatic facial expression analysis techniques and discussed key challenges related to feature extraction and emotion classification. Later, Zeng et al. [1] expanded this area by reviewing audio-visual affect recognition methods and highlighting the importance of combining multiple modalities for detecting spontaneous emotions.

With the growth of deep learning, several approaches have improved the accuracy of recognition systems. Liu et al. [3] introduced a Boosted Deep Belief Network that focuses on iterative feature learning and selection for facial expression recognition. Jung et al. [8] showed that joint fine-tuning of deep neural networks can significantly improve classification performance. Similarly, Hasani and Mahoor [6]

proposed a CNN-CRF model for recognizing facial expressions in video sequences, which helps maintain better temporal consistency in predictions.
Several survey studies have also explored practical challenges in FER systems. Sariyanidi et al. [2] examined different registration methods, representation techniques, and issues caused by pose variations. Li et al. [7] reviewed deep learning- based FER approaches and discussed problems such as dataset limitations, overfitting, and variations due to lighting conditions and head movement. In addition, De la Torre et al.

[5] introduced IntraFace, a framework designed for facial feature tracking and head pose estimation, which is useful for real-time applications.
Although these studies have made strong contributions to facial expression recognition and emotion detection, most of them mainly focus on emotion classification as a separate task. Very few works combine facial expression analysis with structured content-based response evaluation in the context of interview performance. The proposed system attempts to address this gap by integrating facial analysis and content relevance evaluation within a single scoring framework designed specifically for automated mock interview assessment.

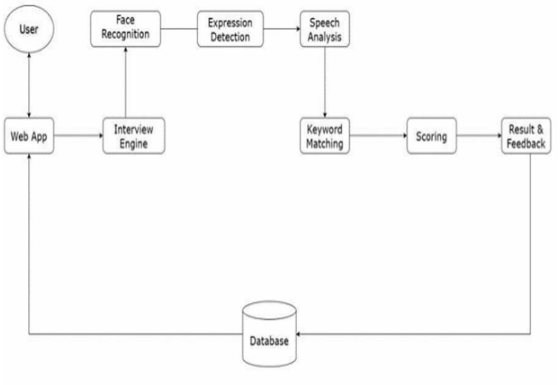
III. SYSTEM DESIGN AND ARCHITECTURE

The mock interview system we've built is a web-based tool that runs structured sessions and evaluates how a candidate performs — both in terms of what they say and how they come across while saying it. It ties together live video and audio capture, facial expression analysis, speech-to-text conversion, and a scoring component under one roof.

A. Overall System Design

When a candidate logs in and starts a session, the system presents them with a set of interview questions and begins recording through their device's camera and microphone. The session ends with a performance report generated automatically from what was captured

The whole thing is built in a modular way — each part of the pipeline does its own job independently. This made it a lot easier to test and fix individual components, and it means the system can grow or change without everything breaking at once. The architecture keeps the user-facing side, data processing, scoring logic, and storage in separate layers.

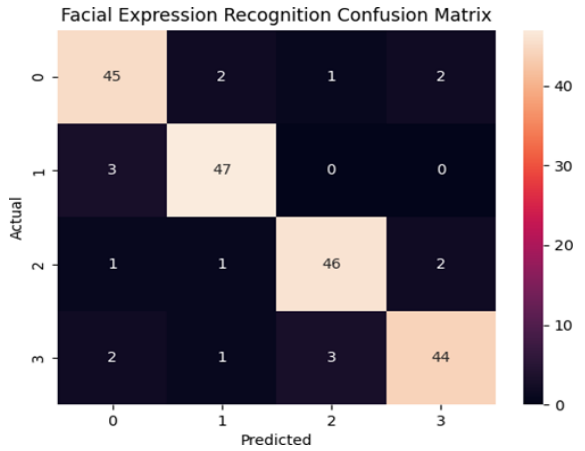


As illustrated in Fig.,the system follows a sequential processing pipeline. The workflow begins with the web application interface, which manages user authentication and session control. Once the interview session starts, the interview engine captures real-time video and audio inputs.

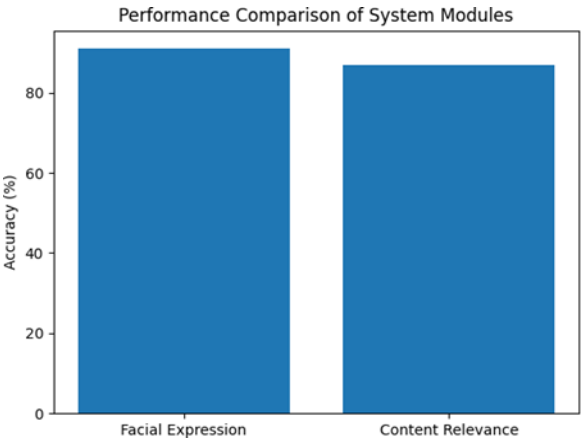
The video stream is processed by the facial recognition module to detect and track the candidate's face. The detected facial data is forwarded to the expression analysis component, which evaluates non- verbal cues such as attentiveness and engagement.

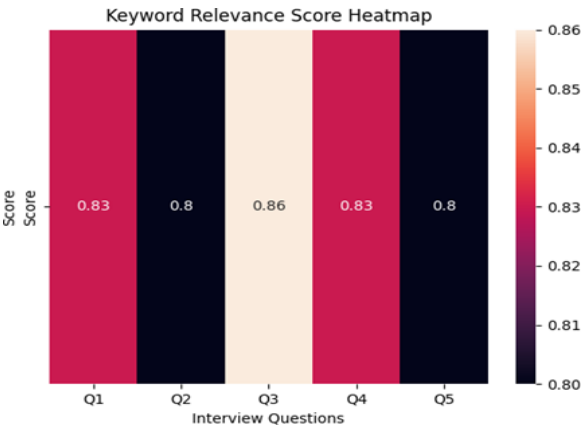
Simultaneously, the captured audio is converted into text using a speech- to-text module. The transcribed response is then analyzed using a keyword matching mechanism to determine the relevance and completeness of the answer.

The outputs from facial expression analysis and content relevance evaluation are passed to the scoring module. A weighted aggregation method is applied to compute the final performance score. Results and feedback are shown to the candidate on their dashboard right after the session wraps up, and all session data gets saved in the database so they can look back at it later.



The following visualizations represent facial expression recognition, performance comparison of system modules & keyword relevance score heatmap:





IV. METHODOLOGY

A. Facial Expression Scoring

Video frames captured during the interview session are processed using a facial detection model to identify facial regions. Each detected frame is classified into predefined expression categories. The system assigns expression scores based on detected engagement levels.

Let E_i represent the expression score of the i^{th}frame

$$F = \frac{1}{n} \sum_{i=1}^{n} E_i \quad (1)$$

The overall facial score F is computed as the average of expression scores across all valid frames where n is the total

number of analyzed frames. This averaging approach reduces noise and ensures stable evaluation across the interview duration.

B. Content Relevance Scoring:

The spoken response is converted into text using a speech recognition module. The transcribed text is compared with a predefined reference answer using keyword matching.

Let: Km = number of matched keywords Kt = total expected keywords

The content relevance score C is computed as:

$$C = \frac{K_m}{K_t} \quad (2)$$

This ratio ensures that the evaluation reflects completeness and relevance of the candidate's response.

C. Final Performance Score:

The final interview performance score S is calculated using a weighted aggregation model:

$$S = w_1 F + w_2 C \quad (3)$$

Where

a) F= facial expression score
b) C= content relevance score
c) w_1 and w_2 are weighting factors such that w_1 + w_2 =1

Combining non-verbal cues with answer quality this way gives a more rounded picture of how the candidate actually performed — not just what they said, but how they came across while saying it

V. IMPLEMENTATION DETAILS

Rather than building a desktop application, we went with a browser-based setup. That way, anyone with a device and internet access can use it — no installs, no setup headaches.

A. Web-Based Interface

We kept the interface as simple as possible. The idea was that a student should be able to open it, log in, start an interview, answer questions, and see their results without having to figure anything out. Minimal friction was the goal.

B. . Facial Analysis Integration

For facial expression detection, a pre-trained model is used. The model processes video frames and classifies expressions. During implementation, it was observed that camera quality and lighting conditions affected performance.

C. Speech Recognition and Text processing

For speech-to-text, we used a tool that works reasonably well under normal conditions. That said, background noise did cause transcription hiccups in some cases, and variations in accent or speaking pace affected the output too. These small errors occasionally threw off the keyword matching and nudged the final score slightly. It's something we'd want to address in a more polished version.

D. Scoring and Report Generation

Once both scores are ready, the scoring module combines them using the weighted aggregation formula to generate a final performance score. The resulting report — which includes the overall score, facial expression score, content relevance score, and some written feedback — appears on the candidate's dashboard as soon as the session is over.

E. Data storage and session management

All interview session data, including responses, computed scores and generated reports are stored securely in the database. This enables candidates to review previous attempts and track performance over time. Session logs are maintained separately to ensure organized data management and future scalability.

VI. RESULTS AND PERFORMANCE

The performance of the proposed system was evaluated by assessing individual modules, including facial expression recognition and content relevance evaluation. Pre-trained models were used for facial analysis and speech recognition, while a custom dataset was developed for interview question and keyword-based answer matching.

i. Experimental setup:

The facial expression recognition module utilizes a pre- trained convolutional neural network (CNN) model trained on standard facial emotion datasets such as FER2013. The model was integrated into the system to classify facial expressions during interview sessions.

Speech recognition was implemented using a pre-trained natural language processing model for speech- to-text conversion. The transcribed responses were processed further for content evaluation.

For content relevance assessment, a custom dataset of interview questions and reference answers was developed. The dataset includes domain-specific technical and behavioral interview questions. Each question is associated with predefined keywords used for evaluating response completeness and relevance.

The evaluation was conducted using multiple mock interview sessions. A total of 60 interview responses were tested for content relevance analysis, while facial expression classification was evaluated using sample video inputs across different candidates.

VII. CONTRIBUTIONS

The proposed AI-based Mock Interview Assistant makes both technical and practical contributions toward automated interview evaluation systems.

A. Technical Contributions

1. Multimodal Interview Evaluation Framework:

Rather than relying on just one signal, the system pulls together facial expression data and content-based scoring into a single unified framework. This makes the evaluation more objective compared to traditional mock interviews that depend entirely on human judgment.

2. Weighted Performance Scoring Model:
We developed a scoring mechanism that blends non-verbal indicators with answer relevance using adjustable weights. This keeps the evaluation balanced and means the reasoning behind any score is transparent and traceable.

3. Custom Interview Dataset for Content Evaluation:
We put together a dataset of interview questions and reference answers specifically for this system. It's designed to support keyword-based scoring and can be adapted for different domains.

4. Real-Time Web-Based Implementation:
Because it runs in the browser, the system works on most devices without any special hardware. It captures and processes both video and audio in real time.

B. Practical Contributions
1.Objective Self-Assessment Tool: Candidates can practice on their own schedule and get structured feedback that doesn't depend on whether someone else is available or in a good mood.
2.Performance Tracking Across Sessions: Session data is saved so that candidates can look back, compare attempts, and see whether they're actually making progress over time.
3.Scalable Interview Preparation Platform: The architecture was designed to be extensible — it can be adapted to different subject areas, making it potentially useful for both academic and professional training contexts.

VIII. FUTURE SCOPE

The proposed AI-based Mock Interview Assistant can be further enhanced through several technical and functional improvements.
Future work may include the integration of advanced deep learning models for improved facial expression classification under varying lighting conditions and camera angles. Incorporating larger and more diverse datasets can improve model robustness and generalization across different user profiles.

The content evaluation module can be enhanced by replacing basic keyword matching with semantic similarity techniques using advanced natural language processing models. This would allow the system to better understand contextual meaning rather than relying only on keyword presence.
Another potential improvement is the inclusion of additional non-verbal indicators such as eye-gaze tracking, head pose estimation, and speech prosody analysis. These enhancements would provide a more comprehensive evaluation of interview performance.
The system can also be extended to support domain-specific interview preparation modules for technical, managerial, and aptitude-based interviews. Integration with cloud-based analytics can further enable large-scale deployment in academic institutions and training platforms.

Overall, the architecture is scalable and can be expanded to incorporate more intelligent evaluation mechanisms while maintaining accessibility and ease of use.

IX. RESULT

This paper described an AI-based mock interview system that evaluates candidates across two dimensions — how they express themselves non-verbally and how relevant their spoken answers are. The system uses a pre-trained model for facial analysis and a speech-to-text pipeline for content scoring, with a custom dataset supporting the keyword-matching component.
Testing across multiple mock sessions showed that the system produces consistent results and gives candidates something genuinely useful to work with. It's not meant to be a substitute for real interviews, but it fills a real gap for students who want structured feedback without having to depend on others.
The underlying architecture is modular and extensible, which means it can grow to support more sophisticated evaluation methods down the line without needing a complete redesign.

REFERENCES

[1] Z. Zeng, M. Pantic, G. I. Roisman, and T. S. Huang,"A survey of affect recognition methods: Audio, visual, and spontaneous expressions," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 31, no. 1, pp. 39–58, Jan. 2009.

[2] E. Sariyanidi, H. Gunes, and A. Cavallaro, "Automatic analysis of facial affect: A survey of registration, representation, and recognition," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 37, no. 6, pp. 1113–1133, Jun. 2015.

[3] P. Liu, S. Han, Z. Meng, and Y. Tong, "Facial expression recognition via a boosted deep belief network," in Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR), 2014, pp. 1805–1812.

[4] M. Pantic and L. J. M. Rothkrantz, "Automatic analysis of facial expressions: The state of the art," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 22, no. 12, pp. 1424–1445, Dec. 2000.

[5] F. De la Torre, W.-S. Chu, X. Xiong, F. Vicente, X. Ding, and J. F. Cohn, "Intraface," in Proc. IEEE Int. Conf. Automatic Face and Gesture Recognition (FG), 2015.

[6] B. Hasani and M. H. Mahoor, "Spatio-temporal facial expression recognition using convolutional neural networks and conditional random fields," in Proc. 12th IEEE Int. Conf. Automatic Face & Gesture Recognition (FG), 2017, pp. 790– 795.

[7] S. Li and W. Deng, "Deep facial expression recognition: A survey," IEEE Transactions on Affective Computing, vol. 13, no. 3, pp. 1195–1215, Jul.–Sep. 2022.

[8] H. Jung, S. Lee, J. Yim, S. Park, and J. Kim, "Joint fine- tuning in deep neural networks for facial expression recognition," in Proc. IEEE Int. Conf. Computer Vision (ICCV), 2015, pp. 2983–2991.