Visionary-Voice: AI-Powered Interview Preparation and Analysis Tool

OPEN ACCESS

Volume: 13

Special Issue: 2

Month: January

Year: 2026

E-ISSN: 2582-0397

P-ISSN: 2321-788X

Citation:
Das, Prabal Deep, and Samay Satyawan Jaunjale. "Visionary-Voice: AI-Powered Interview Preparation and Analysis Tool." *Shanlax International Journal of Arts, Science and Humanities,* vol. 13, no. 2, 2026, pp. 151–57.

DOI:
https://doi.org/10.34293/sijash.v13iS2-i1-Jan.10473

Prabal Deep Das
Assistant Professor, Department of Information Technology and Data Science
Vidyalankar School of Information Technology, Mumbai, Maharashtra, India

Samay Satyawan Jaunjale
Department of Information Technology and Data Science
Vidyalankar School of Information Technology, Mumbai, Maharashtra, India

Abstract
The paper proposes an AI-based multimodal analysis system which can provide an aid to the individual in improving the formal communication skills, in gaining self-confidence and control in non-verbal expressions among students and jobseekers. The proposed system uses various modules like resume-based question generation, speech-to-text conversion and computer vision-based facial analysis, to evaluate verbal clarity, pacing, and other communication-based parameters by processing the recorded audio data through transform based-models for more precise translation and proficiency scoring. The recorded video frames are examined through convolution neural networks to evaluate visual presence and behavioural clues. The outputs from all sections are collated to create a detailed and customized feedback report for focusing on strengths, weaknesses and probable ways of progress. The data-based testing indicates that the system can identify the communication gaps and provide significant guidance, which can assist the user in enhancing both verbal as well as non-verbal communications required for interview preparedness. Instead, the proposed system is scalable and serves as an unbiased and skill-building chance for people in academic and career development processes.

Keywords: Artificial Intelligence, Multimodal Analysis, Interview Performance Evaluation, Automated Feedback System, Speech Analysis, Facial Expression Recognition

Introduction

Interview performance in contemporary recruitment has grown to be a multi-dimensional assessment platform that decidedly has gone way beyond technical ability alone. Employers are increasingly examining candidates on how well they can communicate, structure their thoughts, regulate verbal pacing, and exhibit confidence through non-verbal displays of eye contact and facial expression [1]. Despite possessing strong knowledge of their subject matter, many students and early-career professionals face significant interview issues due to excessive fillers, poor enunciation, monotony, irregular eye contact, and lack of facial expression [2]. These communication challenges more often than not reduce the net impression of capability and thus thin out the chances of success.

Traditional interview preparation techniques, such as mock interviews among peers, instructor mentoring, or coaching academies, have limited consistency, personalization, and scalability. Human evaluators innately

provide subjective judgment of feedback based on one's judgment, experience, and skill of observation [3]. Aside from that, quality mentorship access is not uniformly distributed, which creates unequal disparities in interview readiness among candidates.

In this aspect, this paper presents an AI-enabled multimodal interview feedback system, which could analyze the verbal and non-verbal traits of a candidate in a more objective and data-driven approach. The complementary modules proposed consist of a multimodal question generator based on resumes [5], transformer-based speech-to-text transcription for verbal performance metric extraction, and CNN models for facial expression, engagement, and eye contact analysis [6]. Together, these modules ensure that speech clarity, the frequency of fillers, pacing, behavior of pausing, facial expressiveness, and visual presence are all comprehensively assessed.

All the extracted insights are summarized into an actionable feedback report pointing out strengths, weaknesses, and recommendations for improvement. The structured personalized feedback provided allows learners to iteratively refine their communicative style, build up their confidence, and enhance adaptability to interviews [7].

The integration of state-of-the-art machine learning techniques with a user-centric interface is an effort by the proposed system to offer an accessible, unbiased, and scalable means of interview skill development. The present work talks about the system design, methodology, performance evaluation, and implications of the system for educational institutions, training centers, and job seekers.

Background

Communication skills are slowly becoming a determining factor for the success of an interview in modern recruitment practices. Employers value expression, reasoning, confidence, and non-verbal communication more than ever before. Sankla and Risbud strongly state that the capability for effective communication, both oral as well as non-verbal, plays a crucial role in determining the impression made on the interviewer, which has a direct effect on the selection of the candidate [1]. Communication skills play a pivotal role in this context.

Despite all these, it is quite apparent that most candidates and young professionals find it difficult to showcase their communication skills during their job interviews. According to a systematic review conducted by Chishiba and Mukuka, candidates may struggle with public speaking fear, expressing organized ideas clearly, and confidence during official communication sessions [2]. Furthermore, most trends and observations in social behaviours indicate a decline in communication skills among young generations attributed by less physically interacting and being highly technology-dependent, as presented in research explaining the decline in communication skills among young generations [8].

Another important consideration in the delivery of an interview is the subjectivity involved in the human judgment. As is clear from the study by Radbruch and Schiprowski, sequencing effects, cognitive distortions, and contrast effects greatly influence the judges' assessment, thereby bringing inconsistencies even in those cases when the aspirants have equal merits [3]. Subjectivities might hamper those aspirants who are already burdened by communications apprehension and lack abilities to practice interviewing.

Inequities in access to training materials for interviews further make these issues even more challenging. Findings by Osterman, on issues related to unequal training, reveal that certain sections of society are given far fewer opportunities to gain skills; as a result, the disparity in work readiness and performance increases [4]. The continuously emerging issue related to communication barriers in international students, owing to variations in language, culture, and contexts, again shows that accessible training assistance in interview processes, with the aid of technology, is the requirement [9].

However, with the latest advancements in Artificial Intelligence, these processes can be dealt with in a data-driven manner as well. The use of text and speech recognition and transformation approaches, such as the transformer networks examined by Deshmukh and Raut, enable effective automation in deriving the

processed and systematic details needed in question generation and relevance mapping as per personalized requirements [5]. Conversely, in the final stages of employee selection, computer vision approaches in deciphering facial expressions, as proposed by Pratama et al. using CNN, can offer an in-depth understanding of emotional displays, engagement, and non-verbal communication in interviews [6]. These advances allow for possibilities of multi-modal analysis in simulating or extending beyond human-level observation.

Finally, AI-assisted personalized feedback tools have shown their efficacy in motivating individual students, developing their communication skills, as well as monitoring their performance. This has been supported by the systematic review done by Durak and Onan, which aptly states the support for self-regulated learning, confidence development, and skill development provided through AI-assisted personalized feedback tools [7]. This further supports various other research works done, stating that AI-assisted personalized feedback tools have shown their efficacy in communicating valuable feedback in fields involving intensive communication [7].

Together, these needs and problems lead to the importance of a combined approach which would be able to deal with the verbal, and the non-verbal aspects of doing an interview. Additionally, the proposed system aims at catering this need through the utilization of NLP techniques, speech recognition capabilities, and CNN analysis of the video for providing learners an unbiased and personal environment for conducting interviews.

Methodology

This section covers the multimodal pipeline of the proposed system, including resume parsing and question generation, audio transcription and verbal analytics, video engagement analysis, and integrated feedback generation. The end-to-end workflow orchestration, evaluation metrics, and concise implementation notes are also covered.

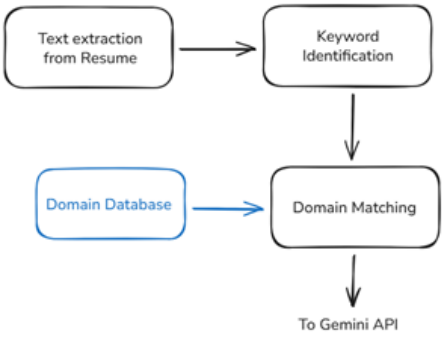

Figure 1 Resume Processing and interview Question Generation

Resume Processing and Interview Question Generation

Figure 1 shows the first step of the proposed workflow, where the system understands and interprets the candidate's resume. It depicts an understanding on the process this uploaded resume goes through for extraction, segmentation, and finally semantic filtering. The graphic points out the employment sections, among others like skills, education, experience, projects, which are extracted into some form of data for the generation of the questions. Certainly, the tenor of the graphic illustration indicates that personal sets of interview questions do come out of the blues, but personal sets of the interview questions must systematically be aligned from the resume categories to the requirements for the candidate when the interview takes place in the particular domain. These sets of blocks represent a process of gradually narrowing down from the text form to filtered concepts finally to tailored direct questions for the purpose of the assessment of the candidate's background.

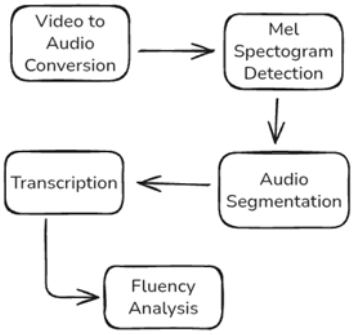

Figure 2 Audio Transcription and Analysis Workflow

Audio Transcription and Analysis Workflow

Figure 2 above highlights the processing of raw audio signals extracted from the candidate's video answer to provide a wide range of verbal assessment measures. This not only highlights how the audio segment can be extracted and cleaned before undergoing automatic speech processing to provide a verbal representation of the answer rendered by candidates. This not only highlights accuracy in various pipeline steps to assess answers by detecting filler words, words per minute calculation, assessing pauses, and evaluating fluency measures but also highlights a technically complex process by adding complexity through each block of processing added to raw audio signals to provide linguistic insights to candidates on improving verbal aptitude.

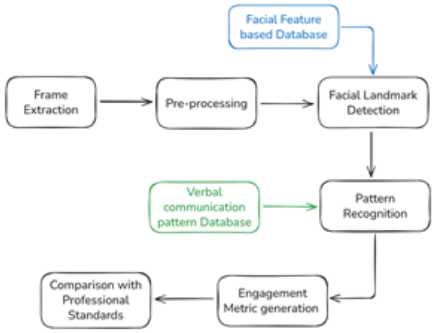

Figure 3 AI-Driven Interview Performace Analysis Workflow

AI-Driven Interview Performance Analysis Workflow

Figure 3 presents the visual analysis processing chain in greater detail, illustrating how the proposed system will analyze nonverbal communication signals by examining the candidate's facial engagement during the interview. Every block in this figure relates to processing video frames from face detection to estimation of gaze direction and expressiveness analysis. This figure illustrates how the system will monitor the candidate's eye contact, measure the candidate's facial expressiveness, and determine the candidate's overall appearance and engagement. This visual analysis processing chain emphasizes the importance of fine non-behavioral signals that indicate how slight movement of the face or shifting of the eyes can alter the perception of confidence. From Figure 3, the level of visual behavior analysis is clear in its depiction that this system will be more than just another face detection system since it will analyze a number of dynamic emotional-positional variables that are part of the interviewer's perception.

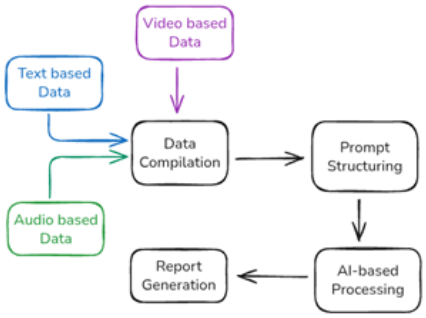

Figure 4 AI-Driven Interview performace Analysis Workflow

Integrated Multimodal Feedback Generation

Figure 4 integrates the findings from the resume, audio, and video modules to illustrate the composition of a candidate evaluation through an extensive analysis. The figure provides an insight into a central performance analysis engine where the streams of data merge in an intelligent multimodal system. The system provides an overall score through the use of basic criteria measuring the speed, fillers, eye contact, expressiveness, and domain performance. The last components provide an insight into the generation of an analytical feedback document with an emphasis on the interpretation of basic observations into digestible and actionable advice. The figure provides an emphasis in relation to the entire system not requiring the analysis of the data but its interpretation for actionable candidate insights.

Result

The methodology was implemented in Python 3.11, using a composite of machine learning, natural language processing, and computer vision libraries to operationalize each part of the pipeline. Key libraries and packages include pdfplumber for resume parsing; OpenCV and Mediapipe for face landmark detection and video frame analysis; Whisper and SpeechRecognition for speech-to-text conversion; pydub and FFmpeg-Python for audio preprocessing; NumPy and Pandas for data management; Scikit-learn for metric calculation; and Matplotlib and Seaborn for visualization during debugging and evaluation. These work in concert to offer a strong foundation for analyzing multimodal meetings.

With the overall processing chain completed, it was noticed that the system produced candidate interview questions appropriately, according to the content of the candidate resume, extracting relevant skills, tools, and projects perfectly. The question generation module was consistently forming candidate interview questions relevant to the contextual background of the candidate, validating that all prior processing phases had successfully extracted and filtered the content from the resume from a semantic viewpoint. In some candidate resumes, the module maintained structural integrity and relevance without failing text extraction and identification.

The transcription and metrics of the audio pipeline is able to perform well regardless of the conditions of the recordings. The speech recognition part of the pipeline is able to provide interpretable transcripts with a low recognition error on a typical speaking pace. The calculated verbal features including Words per Minute, Frequency of Filler Words, Number of Pauses, and Rhythm of Speech were found to vary significantly in the user recordings, thus successfully differentiating the communication styles of the user. Excessive fillers and unpunctual pauses were likewise identified in low-quality answers, proving the sensitivity of the verbal features module of the software.

The analysis of engagement with the video allowed the use of relevant data to interpret the quality of nonverbal communication. During the recorded tests, the system functioned very well to ensure robust processing for the facial and eye regions, and the functionality to distinguish between the two types of frames, namely those that are expressive and those that are not. The engagement measure, which takes into

account eye contact, expressiveness, and cantering, showed a very clear variation between the confidence and hesitations levels of the speakers to ensure that the system can capture the behavioral patterns related to the actual observations made by the interviewer during processing. The system was able to integrate both verbal and non-verbal outputs effectively for significant feedback reports within the multimodal performance analysis module. The reports were able to address strengths like good pace and good eye contact while pointing out areas that require improvement, such as filler elimination, response organization, and visual attention maintenance. Organized reports have been shown to systematically correspond to predefined communication performance criteria and thus confirm responsive functionality for the multimodal fusion model correctly used in the system. Moreover, sensitive analyses were obtained that contained particular and quantifiable goals for improvement and not general statements about overall performance in interviews or public speaking engagement. Overall, the system was able to effectively function with heterogeneous inputs by providing accurate extraction results, precise calculation for metric values, and valid interpretable analyses for improvement purposes. The feasibility and practicality of developing a multimodal interview assessment framework with AI technology assistance for academic and training purposes as well as for skills building platforms was realized through the overall successful functionality and integration of all modules by utilizing Python version 3.10.3 and corresponding modules accordingly.

Conclusion

The complete system of multimodal interview analysis was successfully implemented in Python 3.10.3, which integrated: pdfplumber for the extraction of a resume in structured form; OpenCV and MediaPipe for face landmark detection and engagement analysis; Whisper and SpeechRecognition for speech-to-text processing; ffmpeg-python and pydub to isolate and pre-process audio; and numpy, pandas, and scikit-learn for the computation of metrics that might be aggregated in the results. This integrated pipeline was tested on a variety of resumes and recorded interview responses with respect to robustness, accuracy, and consistency.

It has reliably extracted skills, education, project descriptions, and domain-specific keywords from various resume formats. By doing so, this enabled the generation of three interview questions personalized to each test case, whose content was aligned with the candidate's technical background and project experience. The layout variations were very tolerated by the parsing pipeline; it always structured the outputs without misclassifying sections in a resume.

Excellent performance in transcription quality, generation of metrics, is contributed by the audio-analysis module. Whisper-based transcription yields clear, contextually appropriate transcripts with minimal error for audio samples of conversational English. The system computes words per minute, filler-word frequency, pause density, and speech-rhythm consistency as verbal metrics, enabling the objective differentiation among fluent, moderately fluent, and hesitant speakers. Verbal issues such as excessive filler usage, too-long pauses, and inconsistent pacing were effectively identified by the module across test responses.

The video engagement analytics component functioned properly under varying light conditions and camera types. The facial detection was stable while tracking ratios for eye contact and centering, visibility of face, and expressiveness variations thereof. Further, their corresponding engagement score was properly correlated to their respective confidence levels; for instance, a forward-looking candidate with expressiveness in his face scored higher than a distracted or expressionless candidate. Further, it correctly tracked the biography behaviours like transient gaze switches or slouching posture to make way for a strong definition for unobtrusive assessment.

Finally, the multimodal fusion engine combined and provided a formatted and interpretable feedback report summarizing strengths and weaknesses along with improvement mechanisms for both verbal and unobtrusive cues. Reports were consistent and accurately presented with respect to conventional interview evaluation criteria and compasses of reports for both verbal and unobtrusive cues were accurate and directly applicable for improved future performances. Results successfully validate and conclude that the system

was capable and attempted all dependable and objective analyses with scalability with varying inputs and variations.

Future Scope

The present system has a good grounding, but there are a few development alterations that can be made to it for improvement in the domains of accuracy, flexibility, and viability. Among the most important development needs would be the multidimensional multilingual component for each module, which must be able to evaluate the interviews conducted by the regional as well as the global language. This must specify the speech recognition technology to include regional accents, focusing more on filler words according to the region.

The other possible productive extension would include the development of a mobile app or a light-weight cross-platform client that would allow individuals to conduct their interview practices on the go without requiring any extensive setup. The major benefit that would be reaped in this consideration would include the fact that not only would it benefit a huge student and job-seeking populace that would not have ready access to laptops or high-performance computing machines, but it would also enable its availability on a reduced setup requirement basis. Future releases would allow functionality like badges, streaks, role-specific scenarios, and leader boards.

In order to further upgrade the video analysis task, some new state-of-the-art emotion recognition techniques using heart mechanics can be incorporated. Further improvement of the model employed in visual analytics using the concept of attention is also expected to increase the accuracy of detection related to levels and/or micro-expression changes. Apart from that, measurement of semantic coherence, assessment of the structure of the answer, and the relevance of context can also be introduced.

This can cause a massive boost in the acquisition of skills in the long run, once introduced in the system: the path for each individual to learn, in which the system traces a set of interviews tried and changes the feedback for the user's improvement pattern. Looking ahead, a pipeline for the extension of the dataset contributed by anonymous users should be introduced as well. The future work will be to scale the system into a holistic interview training life cycle that learns by behaviours, supports multi-languages and multi-domains, and provides a mechanism for fine-grained adaptive and engaging feedback for various career readiness use cases.

References

1. Sankla, D. R., & Risbud, M. M. (2025). Mastering interview skills: Techniques and guidelines to crack interview. *InSight Bulletin, 2*(3), 75–78.
2. Chishiba, G., & Mukuka, J. (2024). Communication skills challenges experienced by first-year university students: A systematic review. *JESBS, 37*(6).
3. Radbruch, J., & Schiprowski, A. (2025). Interview sequences and the formation of subjective assessments. *Rev. Econ. Stud., 92*(2), 1226–1256.
4. Osterman, P. (2020). Skill training for adults: New research on training disparities. MIT Sloan IWER.
5. Deshmukh, A., & Raut, A. (2025). Applying BERT-based NLP for automated resume screening. *Annals of Data Science, 12,* 591–603.
6. Pratama, M. R., et al. (2025). Student expression detection based on facial image using CNN. *Ceddi Journal of Education, 4*(1).
7. Durak, H. Y., & Onan, A. (2025). A systematic review of AI-based feedback in educational settings. *J. Comput. Social Science, 8,* art. 96.
8. Konrad, R., & Abrahams, M. (2025). Why young people are struggling to communicate. TIME Magazine.
9. Moghaddam, M. M. (2024). International students' communication challenges in higher education. *European Journal of Psychology of Education, 39,* 4617–4646.