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.2026.153114

SkillMirror: An AI-Based Communication and Interview Training Platform Using Speech and Computer Vision Analysis

Shreyas S. Gadekar[1], Atharv N. Mane[2], Prasad T. Kalavikattekar[3], Aniket M. Kamble[4], Karan P. Ghatage[5], Prof. Avani G Shahane[6]

DKTE's Textile and Engineering Institute, Ichalkaranji Sangli, India[1-5]

Guide, DKTE's Textile and Engineering Institute, Ichalkaranji Sangli, India[6]

Abstract: Effective communication and interview skills are essential for student employability, yet traditional training methods often lack personalization and real-time feedback. This paper presents SkillMirror, an AI-based communication and interview training platform that enhances both verbal and non-verbal skills through a multimodal analysis approach. The system simulates interview scenarios where user responses are captured via audio and video. Speech-to-text and natural language processing techniques are employed to evaluate grammar, fluency, and linguistic coherence, while computer vision methods analyze non-verbal cues such as facial expressions, posture, and eye contact. Based on these analyses, the platform provides real-time feedback, performance metrics, and personalized improvement suggestions.

Additionally, the system includes a quiz module for knowledge assessment and a feedback module for progress tracking and motivation. By integrating speech processing and behavioral analysis, SkillMirror offers a scalable and intelligent solution for improving communication skills and interview readiness.

Keywords: Artificial Intelligence, Interview Training, Speech-to-Text, Natural Language Processing, Computer Vision, Multimodal Analysis, Communication Skills, Real Time FeedbackVision, Multimodal Analysis, Communication Skills, Real-Time Feedback

I. INTRODUCTION

In recent years, effective communication and interview performance have emerged as critical competencies for students and job seekers in an increasingly competitive professional environment. While technical knowledge remains essential, the ability to articulate ideas clearly, demonstrate confidence, and exhibit appropriate non-verbal behavior significantly influences hiring decisions. However, many students face challenges such as lack of practice, communication anxiety, and limited access to personalized training resources, which hinder their overall performance in real-world interview scenarios.

Traditional methods of communication training, including classroom-based learning and mock interviews, often lack scalability, consistency, and objective evaluation. These approaches typically depend on human evaluators, making them time-consuming, subjective, and inaccessible to a large number of learners. As a result, there is a growing need for intelligent systems that can provide automated, real-time, and personalized feedback to improve both verbal and non-verbal communication skills.

Recent advancements in Artificial Intelligence (AI), particularly in Natural Language Processing (NLP) and Computer Vision (CV), have enabled the development of systems capable of analyzing human speech and behavior with high accuracy. Speech processing techniques, such as speech-to-text conversion, allow the extraction and analysis of linguistic features including grammar, fluency, and coherence. Simultaneously, computer vision methods facilitate the detection and interpretation of non-verbal cues such as facial expressions, posture, gestures, and eye contact, which are essential components of effective communication.

In this context, this paper proposes SkillMirror, an AI-based communication and interview training platform that integrates speech processing, natural language understanding, and computer vision to deliver a comprehensive and multimodal evaluation of user performance. The system provides a simulated interview environment where users can respond to both technical and non-technical questions while being continuously analyzed through audio and video inputs.

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.2026.153114

Based on this analysis, the platform generates real-time feedback, performance metrics, and personalized recommendations for improvement.

The primary objective of SkillMirror is to bridge the gap between theoretical knowledge and practical communication skills by offering an accessible, scalable, and intelligent training solution. By leveraging multimodal AI techniques, the system aims to enhance user confidence, improve communication effectiveness, and increase overall employability.

II. LITERATURE REVIEW

Recent advancements in Artificial Intelligence (AI) have enabled the development of intelligent systems capable of analysing both verbal and non-verbal aspects of human communication. This section reviews key technologies relevant to the proposed system, including speech processing, natural language processing, and computer vision-based behavioural analysis.

2.1 Speech-to-Text and Speech Processing

Speech recognition systems have evolved significantly with the adoption of deep learning and transformer-based architectures. These systems are capable of converting spoken language into text with high accuracy across diverse environments. Large-scale supervised models trained on extensive datasets have demonstrated strong generalization capabilities and robustness in real-world scenarios. Such systems enable further linguistic analysis, including grammar evaluation and fluency assessment.

Key Contributions:

- Enables real-time conversion of speech into textual data
- Supports downstream NLP tasks such as grammar and fluency analysis
- Demonstrates robustness across multilingual and noisy environments [1]

2.2 Natural Language Processing (NLP)

Natural Language Processing techniques are widely used to analyse textual data for syntactic and semantic correctness. NLP models can detect grammatical errors, evaluate sentence structure, and assess coherence, making them suitable for automated communication assessment systems.

Key Contributions:

- Grammar and syntax error detection
- Evaluation of linguistic coherence and sentence complexity
- Enables automated feedback generation for spoken responses

2.3 Computer Vision for Behavioural Analysis

Computer vision frameworks have enabled real-time analysis of human behavior through visual data. MediaPipe provides a modular pipeline for processing multimodal data, allowing efficient detection of facial landmarks, hand gestures, and body posture. Its graph-based architecture ensures low-latency processing suitable for interactive applications.

OpenCV, on the other hand, offers a comprehensive set of tools for face detection, tracking, and motion analysis. However, studies indicate that real-time tracking systems face challenges such as illumination variation, occlusion, and dynamic backgrounds, which can affect accuracy.

Key Contributions:

- Real-time detection of facial expressions and gestures [2]
- Posture and eye-contact analysis using landmark detection
- Widely adopted frameworks for interactive vision-based systems

2.4 Facial Expression and Emotion Recognition

Facial expression recognition systems leverage deep learning models trained on large-scale datasets to classify human emotions. Datasets such as AffectNet have enabled the development of models capable of recognizing a wide range of emotional states, contributing to behavioral and psychological analysis.

Key Contributions:

- Detection of emotional states from facial features
- Supports analysis of confidence and engagement levels
- Enhances non-verbal communication assessment

2.5 Research Gap

Despite significant advancements in individual domains, existing systems exhibit several limitations:

- Most systems focus on either speech analysis or visual analysis, but not both
- Lack of integrated multimodal frameworks combining NLP and computer vision
- Limited availability of real-time feedback systems for interview training

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.2026.153114

- Absence of personalized performance tracking and improvement suggestions
- High dependency on human evaluators in traditional systems

2.6 Motivation for Proposed System

To address these limitations, the proposed system integrates speech processing, natural language analysis, and computer vision into a unified platform. This multimodal approach enables comprehensive evaluation of both verbal and non-verbal communication, along with real-time feedback and performance analytics.

III. PROPOSED SYSTEM

This section presents the design and architecture of SkillMirror, an AI-based communication and interview training platform that integrates speech processing, natural language processing, and computer vision to evaluate both verbal and non-verbal communication skills.

3.1 System Overview

The proposed system is designed as an interactive web-based platform that simulates real-world interview scenarios. It consists of three primary modules: (i) Quiz Module, (ii) Chatbot Interview Module, and (iii) Feedback Module. The system captures user responses through audio and video inputs and processes them using AI-based models to generate real-time feedback.

The workflow of the system begins with user authentication, followed by access to the dashboard, where users can select different modules. During the interview session, the system continuously analyses user input and updates performance metrics dynamically.

3.2 System Architecture

The overall architecture of the system follows a multimodal processing pipeline, integrating multiple components as shown below:

Input Layer:
- Audio input from microphone
- Video input from webcam

Processing Layer:
- Speech-to-text conversion module
- Natural language processing module
- Computer vision module

Analysis Layer:
- Grammar and fluency evaluation
- Gesture, posture, and eye-contact detection
- Emotion and behavioural analysis

Output Layer:
- Real-time feedback generation
- Performance scoring
- Suggestions for improvement

This layered architecture ensures modularity, scalability, and efficient real-time processing of multimodal data.

3.3 Quiz Module

The Quiz Module is designed to assess the user's theoretical knowledge through structured questionnaires. It consists of two categories:

- **Technical Questions:** Domain-specific questions related to programming, data structures, and core subjects
- **Non-Technical Questions:** Behavioural and HR-related questions

This module helps users prepare for interviews by strengthening their conceptual understanding and improving response formulation.

3.4 Chatbot Interview Module

The Chatbot Interview Module simulates a real-time interview environment. The system requests access to the user's camera and microphone to capture multimodal inputs.

Working Process:

1. The system presents interview questions sequentially
2. The user responds verbally

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.2026.153114

3. Audio is captured and converted into text using speech recognition
4. Video frames are processed simultaneously for behavioural analysis

This module ensures an interactive and realistic interview experience, enabling users to practice under simulated conditions.

3.5 Speech and NLP Processing Module

The speech processing module converts spoken responses into textual data using speech-to-text techniques. The generated text is then analysed using Natural Language Processing (NLP) methods.

Key Functionalities:

- Grammar error detection
- Fluency and speech coherence analysis
- Sentence structure evaluation

These features allow the system to assess the linguistic quality of user responses and identify areas for improvement.

3.6 Computer Vision Module

The computer vision module analyses video input to evaluate non-verbal communication. It processes facial and body features to extract behavioural cues.

Key Functionalities:

- Face detection and facial landmark extraction
- Eye-contact tracking
- Posture and gesture analysis
- Basic emotion recognition

This module plays a crucial role in assessing confidence, engagement, and overall presentation skills.

3.7 Feedback and Performance Analysis Module

The Feedback Module integrates outputs from both NLP and computer vision modules to generate comprehensive performance reports.

Key Features:

- Real-time feedback during interview sessions
- Performance scoring based on multiple parameters
- Identification of strengths and weaknesses
- Personalized improvement suggestions
- Badge-based reward system for motivation

The feedback is presented in a structured format, enabling users to track their progress over time.

3.8 System Workflow Summary

The complete workflow of the system can be summarized as follows:

1. User logs into the system
2. Selects quiz or interview module
3. Provides responses via audio and video
4. System processes input using AI models
5. Feedback is generated and displayed
6. Performance data is stored for future analysis

The proposed system leverages a multimodal AI framework to provide a comprehensive and scalable solution for interview training. By integrating speech analysis, natural language understanding, and computer vision, SkillMirror enables holistic evaluation of communication skills, thereby improving user confidence and employability.

IV. METHODOLOGY

This section describes the methodology used in the proposed SkillMirror system, focusing on data acquisition, processing pipeline, and multimodal analysis of user responses. The system follows a structured workflow to process audio and video inputs and generate meaningful feedback.

4.1 Data Acquisition

The system collects multimodal input data from the user during the interview session:

- **Audio Data:** Captured through the microphone for speech analysis

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.2026.153114

- **Video Data:** Captured through the webcam for behavioral analysis

These inputs are processed in real-time to ensure immediate feedback.

4.2 Data Processing Pipeline

The methodology follows a sequential pipeline consisting of three main stages:

1. **Pre-processing Stage**
   - Audio signals are cleaned to reduce noise
   - Video frames are extracted from the live stream
   - Input data is normalized for consistent processing
2. **Feature Extraction Stage**
   - Speech features are extracted from audio signals
   - Textual features are derived after speech-to-text conversion
   - Visual features such as facial landmarks and body posture are detected
3. **Analysis Stage**
   - Linguistic analysis using NLP
   - Behavioural analysis using computer vision
   - Integration of results for feedback generation

4.3 Speech Processing and NLP Analysis

The audio input is converted into text using speech recognition techniques. The resulting text is processed using Natural Language Processing methods.

Techniques Used:

- Tokenization and sentence segmentation
- Grammar and syntax evaluation
- Fluency and coherence analysis

The system evaluates:

- Grammatical correctness
- Sentence structure
- Clarity of expression

4.4 Computer Vision-Based Analysis

The video input is analysed using computer vision techniques to extract non-verbal cues.

Techniques Used:

- Face detection and landmark extraction
- Eye gaze tracking for eye-contact analysis
- Pose estimation for posture detection
- Gesture recognition

These features help assess:

- Confidence level
- Engagement
- Body language

4.5 Multimodal Data Fusion

The outputs from the NLP and computer vision modules are combined to generate a unified assessment.

- Verbal features (speech + text)
- Non-verbal features (video analysis)

This fusion enables a comprehensive evaluation of communication skills.

This section describes the methodology adopted in the SkillMirror system for analyzing user communication through multimodal data. The system processes audio and video inputs in real time to evaluate both verbal and non-verbal aspects of communication.

4.6 Feedback Generation

Based on the analysed data, the system generates:

- Real-time feedback
- Performance scores
- Personalized suggestions

The feedback is structured to highlight strengths and areas for improvement.

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.2026.153114

V. RESULTS AND EXPERIMENTAL DISCUSSION

This section presents the results and experimental analysis of the proposed SkillMirror system. The evaluation focuses on the effectiveness of speech analysis, non-verbal behavior detection, and overall user performance improvement.

5.1 Experimental Setup

The system was tested in a controlled environment using a standard webcam and microphone. Multiple users participated in simulated interview sessions, responding to both technical and non-technical questions. The system processed audio and video inputs in real time and generated feedback based on predefined evaluation parameters.

The performance of the system was analysed based on:

- Accuracy of speech-to-text conversion
- Effectiveness of grammar and fluency analysis
- Reliability of non-verbal cue detection
- User feedback and improvement trends

5.2 Results of Speech and Language Analysis

The speech processing module successfully converted spoken responses into text with high accuracy under normal acoustic conditions. The NLP module effectively identified grammatical errors, sentence structure issues, and fluency gaps.

Observations:

- Minor inaccuracies occurred in noisy environments
- Grammar detection was consistent for simple and moderately complex sentences
- Fluency evaluation helped identify hesitation and repetition patterns
- Overall, the system demonstrated reliable performance in analyzing verbal communication.

5.3 Results of Non-Verbal Behaviour Analysis

The computer vision module was able to detect facial features, eye contact, and posture in real time. The system performed well under adequate lighting conditions and stable camera positioning.

Observations:

- Eye contact detection was effective for frontal face orientation
- Posture analysis identified slouching and improper positioning
- Gesture detection provided additional behavioral insights

However, performance was affected by poor lighting and partial occlusion.

5.4 Integrated Performance Evaluation

The multimodal approach enabled comprehensive evaluation by combining verbal and non-verbal analysis. The feedback module generated detailed performance reports, including scores and improvement suggestions.

Key Outcomes:

- Users received structured feedback on communication skills
- Performance scores reflected both linguistic and behavioural aspects
- The system successfully identified strengths and weaknesses

5.5 User Performance and Feedback

Users reported improved awareness of their communication style after multiple sessions. The feedback system helped users identify areas such as grammar mistakes, lack of eye contact, and poor posture.

Findings:

- Noticeable improvement in user confidence over repeated sessions
- Better sentence formation and reduced grammatical errors
- Improved body language and engagement

5.6 Discussion

The experimental results demonstrate that the proposed system is effective in providing real-time, multimodal feedback for interview training. The integration of speech processing, NLP, and computer vision enables a holistic evaluation of communication skills.

Compared to traditional methods, the system offers:

- Automated and objective assessment
- Real-time feedback
- Scalability and accessibility

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.2026.153114

Despite these advantages, certain limitations such as environmental dependency and model accuracy need to be addressed for further improvement.

VI. CONCLUSION AND FUTURE SCOPE

6.1 Conclusion

This paper presented SkillMirror, an AI-based communication and interview training platform designed to enhance both verbal and non-verbal communication skills through a multimodal analysis approach. The system integrates speech processing, natural language processing, and computer vision techniques to evaluate user performance in simulated interview scenarios.

The proposed platform successfully provides real-time feedback by analyzing grammatical correctness, fluency, and linguistic coherence, along with behavioral aspects such as eye contact, posture, and gestures. The inclusion of quiz-based assessment and performance tracking further strengthens the learning process by enabling continuous improvement.

Experimental observations indicate that the system is effective in identifying communication gaps and providing structured feedback to users. By offering an automated, scalable, and accessible solution, SkillMirror helps bridge the gap between theoretical knowledge and practical interview readiness. The system contributes to improving user confidence, communication effectiveness, and overall employability.

6.2 Future Scope

Although the proposed system demonstrates promising results, several enhancements can be incorporated to improve its performance and applicability:

- Integration of advanced emotion recognition for deeper behavioral analysis
- Support for multiple languages to enhance accessibility
- Improvement in noise handling and robustness of speech recognition
- Deployment of more accurate deep learning models for gesture and posture detection
- Development of adaptive AI interviewers capable of dynamic question generation
- Cloud-based scalability for large-scale user access
- Integration with placement and recruitment platforms

Future advancements in AI and multimodal analysis can further enhance the system's capabilities, making it a more intelligent and comprehensive solution for communication training and interview preparation.

REFERENCES

[1]. A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McCelvey, and I. Stuever, "Robust Speech Recognition via Large-Scale Weak Supervision," in Proceedings of the International Conference on Machine Learning (ICML), 2022.
[2]. Google, "Media Pipe: A Framework for Building Multimodal Applied Machine Learning Pipelines," 2019.
[3]. G. Brad Ski, "The OpenCV Library," Dr. Dobb's Journal of Software Tools, 2000.
[4]. A. Molla Hosseini, B. Hasani, and M. H. Mahoor, "Affect Net: A Database for Facial Expression, Valence, and Arousal Computing in the Wild," IEEE Transactions on Affective Computing, vol. 10, no. 1, pp. 18–31, 2019.

