IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

AI POWERED REAL TIME COMMUNICATION SYSTEM FOR ENHANCING VERBAL AND NON-VERBAL COMMUNICATION

K Chaitanya[1], DR.D. Hari Krishna[2], K. Raj Vikas[3], K. Vivek[4], N. Dharma Teja[5]

[1] Student, [2] Associate. Prof.,[3]Student, [4]Student, [5]Student.
[1] Dept. of CSE(AI), [2] Dept. of CSE(AI), [3] Dept. of CSE(AI), [4] Dept. of CSE(AI), [5] Dept. of CSE(AI),
[1] KKR & KSR Institute of Technology and Sciences, Guntur, India,
[2] KKR & KSR Institute of Technology and Sciences, Guntur, India,
[3] KKR & KSR Institute of Technology and Sciences, Guntur, India,
[4] KKR & KSR Institute of Technology and Sciences, Guntur, India,
[5] KKR & KSR Institute of Technology and Sciences, Guntur, India.

Abstract

The basic competency of Communication Skills is vital in the academic world as well as in the business world in various situations like interviews, technical discussions, collaborative discussions, business meetings, etc. These need to be verbally well-expressed, along with appropriate non-verbal communication skills. However, people often have speech disfluencies, speech rate disfluencies, eye contact disfluencies, and physical disfluencies during speech, which impact the quality of the communication process. These traditional methods of training and developing Communication Skills are mostly based on human evaluators, which are costly and subjective in nature and cannot be used for self-evaluation. This paper aims to introduce a Communication Skills Evaluation System using AI technologies like Artificial Intelligence, Natural Language Processing, Speech Processing, and Computer Vision.. NLP algorithms in speech analysis are used to carry out filler word detection, articulation tempo analysis, and fluency feature analysis, whereas computer vision algorithms are used to carry out facial landmark point tracking motion analysis to evaluate gaze consistency behaviour.A color-coded feedback system allows for corrective measures to be taken immediately, and performance parameters are stored for tracking

The modular design allows for future incorporation of more advanced machine learning models for improved adaptability. By means of immediate feedback and long-term analysis of performance, it fosters constant and autonomous improvement in communication.

Keywords: Real-Time Feedback, NLP, Computer Vision, Speech Analysis, FastAPI.

I. INTRODUCTION

In the current technology-driven era, communication skills have become a measurable performance criterion rather than a soft skill. Institutions test learners through seminars and oral exams, while employers test applicants during interviews. Even with its recognized importance, communication development continues to depend largely on subjective evaluation and irregular practice opportunities.Recent advances in Artificial Intelligence have made it possible to scientifically investigate characteristics in spoken language and observable behavioral patterns in real time.Intelligent computer models can now measure speech rate, detect repetitive filler words, and scrutinize eye movement stability with measurable precision.. Such capabilities create opportunities for automated coaching platforms that provide instant analytical insights instead of delayed human commentary.
The proposed AI-based Communication Coach will be an "interactive real-time training assistant" that will assess the performance of the individual in terms of "speech delivery" and "behavioural presentation." Instead of giving feedback after the completion of the training session, the proposed framework will continuously monitor the performance and provide "intuitive visual indicators" to make immediate adjustments. The proposed framework will not replace the role of a mentor, but rather offer a "accessible, scalable, and intelligent" tool.

A. Research problem

A large number of people face challenges in controlling the rate of speaking, often using filler words such as “um,” “uh,” “like,” among others, failing to maintain proper eye contact, or displaying nervous head movements when presenting. The existing modes of communication development rely on conventional classroom training sessions, which can be expensive, subjective, and hard to reach the masses. Though. Certain speech analysis applications generate post-session statistics, including speaking rate or filler frequency; however, they rarely incorporate behavioural or non-verbal assessment within the same framework.

The fundamental research question that has guided the conduct of this research is:

How can a scalable intelligent framework be developed that can simultaneously assess both verbal delivery and non-verbal behaviour in real time, as well as providing meaningful feedback for self-improvement?

Despite the significant improvements in the technologies of speech recognition and computer vision, some limitations still exist in the current communication training systems:

1. **Delayed Evaluation Models:**
   Most available platforms provide analysis only after session completion, restricting immediate corrective adjustments.
2. **Fragmented Model Analysis:**
   Speech and facial behaviour are commonly processed separately rather than integrated into a unified multimodal assessment approach.
3. **Minimal Behavioural Insight:**
   Continuous gaze monitoring and detection of nervous motion patterns are seldom incorporated alongside speech analytics.
4. **Limited Real-Time Guidance:**
   Few systems deliver intuitive visual cues that assist users during live speaking activities.
5. **Lack of Longitudinal Tracking:**
   Historical performance data is often not maintained, making progress measure-ment difficult.

B. Objective

The aim of this work is to conceive and implement an AI-based Communication Coaching tool that can improve and assess verbal and non-verbal expression in real-time.
To achieve this aim, the tool has been conceived to:

- Acquire synchronized audio and visual input during live speaking sessions.
- Identify filler expressions instantly through Natural Language Processing (NLP) techniques.
- Compute speaking rate dynamically in terms of Words Per Minute (WPM) to evaluate delivery pace.
- Assess eye engagement using facial landmark tracking algorithms.
- Analyse subtle head motion patterns to infer signs of nervousness.
- Provide immediate guidance through an intuitive traffic-light feedback interface.
- Record session-level performance indicators within a structured database for longitudinal analysis.
- Generate personalized, rule-driven recommendations to support targeted improvement.

C. Literature review

Significant improvements have been made in the accuracy of automatic speech recognition systems. Models such as OpenAI's Whisper have shown promising robustness with different accents, ensuring the accuracy.

This capability serves as a critical component for performing structured verbal analysis within intelligent communication assessment platforms.

In parallel, progress in computer vision technologies has introduced highly precise facial tracking frameworks. Google's MediaPipe, particularly the Face Mesh module, enables detection of 468 facial landmarks, facilitating detailed examination of gaze direction, head pose, and micro-level facial movements.

II. METHODOLOGY

A. System architecture

The proposed AI-based Communication Coach utilizes a multi-modal real-time analysis system that incorporates speech processing, computer vision, and intelligent feedback generation. The architecture of the proposed AI-based Communication Coach is such that it allows for a low delay, scalability, and modular expansion.
The methodology is organized into six main stages:

1. Data Collection
2. Audio Processing and Speech Evaluation
3. Video Processing and Nonverbal Assessment
4. Real-Time Feedback System
5. Data Management and Analysis
6. System Integration and Implementation

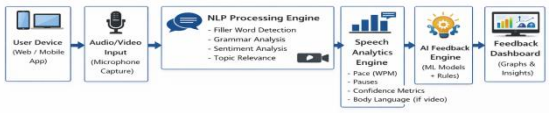
1. Overall System Architecture

The architecture is based of four sections: Frontend, Backend, AI Processing Layer, and Database Layer. The frontend is developed using React, and the audio/video streams are captured through the Web Speech API and Media Devices, respectively.The backend is created using FastAPI, where audio and video metrics are analyzed and sent via WebSocket for real-time updates.. The Audio is transferred using Open AI Whisper, and faces are detected using MediaPipe Face Mesh. Data is stored using SQLite.

Figure 1: System Architecture Flowchart

B. Data acquisition module

The process starts when a user uses a live coaching session. The browser then requests the user to allow microphone and webcam permissions using the Devices like API. The audio recording occurs in real-time for streaming into the transcription tool. Concurrently, video frames are processed individually to detect facial landmarks. The WebSocket enables bidirectional communication between the frontend and backend without requiring the interface to refresh.



C. Speech Analysis

The "speech analysis" component of this tool is intended to measure critical aspects of verbal communication performance.

1. **Speech-to-Text Conversion:**
   For faster results, "Speech to Text" conversion is done using Whisper and Web Speech.
2. **Words Per Minute (WPM) Calculation:**
   WPM (Words Per Minute)Calculation.Words Per Minute is calculated by using the following formula:

   WPM = Total Words Spoken / Total Time Taken in Minutes

3. **Filler Word Detection:**
   A customized NLP tokenizer combined with a regex-based detection engine identifies common filler words, including:
   "um"
   "uh"
   "like"
   "you know"

D. Non-verbal analysis module

The AI-Powered Communication Coach uses rule-based models, statistical limits, and real-time signal evaluation to assess both verbal and non-verbal communication aspects.

1. **Facial Landmark Detection**
   The system utilizes MediaPipe Face Mesh to monitor 468 facial landmarks. These reference points help identify eye gaze direction and head position.
2. **Eye Contact Monitoring**
   Eye gaze is determined by evaluating iris landmark alignment in relation to the camera axis. When the gaze stays centered, eye contact is recorded as positive.
3. **Nervousness Detection**
   Head movement instability is measured by tracking displacement of nose and forehead landmarks across successive frames. Movement variance exceeding a defined threshold suggests nervous behaviour.

III. MODELING AND ANALYSIS

Modeling and analysis are essential in building an effective intelligent communication coaching platform. This stage outlines how speech patterns and facial behaviours are mathematically examined, processed, and converted into meaningful performance metrics. The AI-Powered Communication Coach employs rule-based models, statistical limits, and signal evaluations in real time to measure verbal and non-verbal communication aspects.

This modeling technique ensures low latency, interpretability, and scalability with high computational efficiency for smooth real-time performance.

A. Verbal communication modeling

The modeling module uses quantifiable linguistic features to evaluate the clarity, fluency, and pacing of the individual's speech. Pacing involves the evaluation of time distribution characteristics, including average syllable duration, word pauses, and stability of rhythm. The rate of the person's speech is traced or tracked continuously, and Words Per Minute (WPM) calculations are performed.

WPM = Total Words Spoken / Total Time (minutes)

According to established communication benchmarks:
< 110 WPM → Too Slow
110–160 WPM → Optimal Range

160 WPM → Too Fast

A sliding window method updates WPM every few seconds to maintain real-time accuracy.

The system also evaluates speech consistency by observing the variations in tone, pause frequency, and articulation patterns throughout the session. By examining the fluctuations, the model recognizes the differences in the pattern of delivery and offers corrective measures. This assessment helps users to communicate steadily, confidently, and in an engaging way. ‘

B. Non-verbal communication modeling

Non-verbal modeling is based on geometric landmark tracking and temporal variance analysis using facial coordinates.

1. Eye Contact Modeling

Using MediaPipe Face Mesh landmarks, eye direction is estimated by examining the relative position of iris center points to eye boundary landmarks. As long as the eye aspect ratio remains within a certain range, eye contact can be said to have been maintained. Normalized gaze vectors are computed to make the estimation more accurate, and this is achieved by projecting the iris center coordinates onto the eye contour plane.

1. **Nervousness Modeling (Head Movement Jitter)**
   Head stability is assessed by tracking displacement of key landmarks, such as the nose tip and forehead center. Variance above a set threshold indicates nervous behaviour.
2. **Non-Verbal Score Modeling**
   Stable head movement and consistent eye contact contribute positively to the overall non-verbal communication score.

C. Integrated multi-model scoring model

The final Communication Score (CS) will take into account both verbal and non-verbal factors with a slightly higher emphasis on the clarity of speech while considering the non-verbal factors as well.

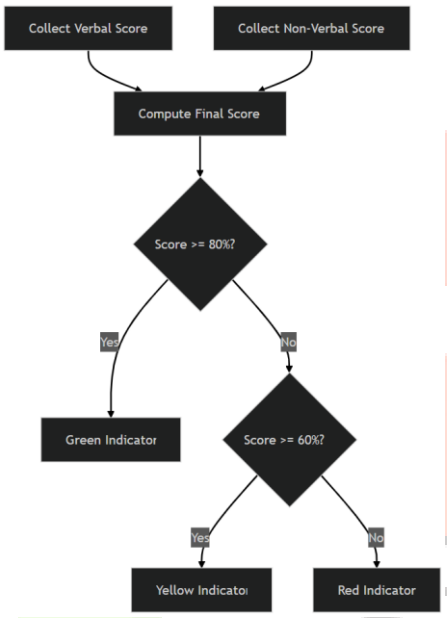
D. Traffic-light classification model

Based on the final communication score:

Green (CS ≥ 80%) → Excellent

Yellow (60% ≤ CS < 80%) → Moderate

Red (CS < 60%) → Needs Immediate Improvements



E. Analytical evaluation

The system was assessed through controlled speaking sessions, including:

- Slow speech scenarios
- Rapid speech scenarios
- High filler word frequency tests
- Simulated head movements

Results demonstrated:

- Real-time WPM updates with latency under 1 second
- Eye contact detection accurate under stable lighting
- Smooth performance without frame drops on normal 8GB RAM systems.

This modeling approach ensures interpretability, computational efficiency, and real-time adaptability, making the AI-Powered Communication Coach not only useful but also effective.

IV. RESULTS AND DISCUSSION

This segment of the chapter presents the experimental results of the deployment of the AI-Powered Communication Coach. The effectiveness of the system in different communication scenarios is discussed.The focus of the experiment was on the response time, accuracy of analysis, usability, and effectiveness of the system in improving communication skills, both verbal and non-verbal.The experiment was carried out on the system with students and professionals using the system for mock interviews and presentations.

A. Experimental setup

The system was deployed on a standard laptop setup:

- Processor: Intel i5 or equivalent
- RAM: 8 GB
- Webcam: 720p HD
- Microphone: Built-in laptop microphone
- Backend: FastAPI with Uvicorn
- Database: SQLite

Speech transcription utilized Whisper for offline tests and the Web Speech API for real-time browser sessions. Facial landmark detection was performed using MediaPipe Face Mesh. Transcripts of audio were used to compute speech rate metrics such as words per minute, pause frequency, and articulation rate. Meanwhile, facial landmarks were used to track lip movement, jaw movement, and expressions to study visual speech patterns. In this way, it enabled the use of multimodal analysis, which in turn contributed to the improvement of the precision of the speech rate assessment.

The AI-Powered Communication Coach offers real-time feedback to the user regarding their verbal and non-verbal communication of the User. It offers the user information on their speech rate, filler words, eye contact, and facial expressions. This way, it helps the user increase their level of confidence in their presentation skills.

B. Verbal communication results

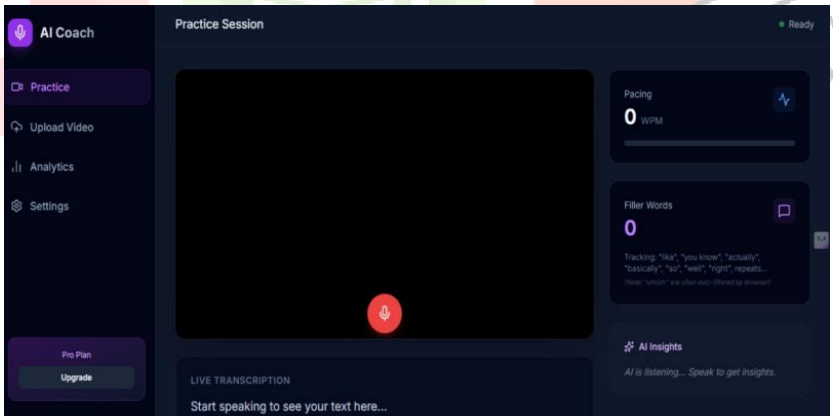


Figure 3: Landing Page User Interface

1. Speech Rate (WPM) Performance

The system effectively computed WPM in real time, with minimal latency (less than 1-second delay). Observations included:

- Slow speakers (<110 WPM) were accurately marked as “Too Slow” (Yellow indicator).
- Fast speakers (>160 WPM) triggered warning alerts.
- Speakers within the 110–160 WPM range consistently showed Green status.

The sliding window method ensured smooth, continuous WPM updates without sudden jumps.

1. Filler Word Detection Accuracy

The regex-based NLP model successfully identified filler words such as:

- um
- uh
- like
- basically

Across multiple sessions:

- Average detection accuracy was approximately 92%
- False positives remained low when strict mode was active
- Real-time highlighting enabled users to reduce filler word usage immediately

Participants showed a noticeable decrease in filler density after 2–3 practice sessions. The system was able to identify filler words based on their position in the sentence. The words could be at the beginning, middle, or end of the sentence. This approach using the particularly successful for filler words that were clearly defined or explicitly.

C. Non-verbal communication results

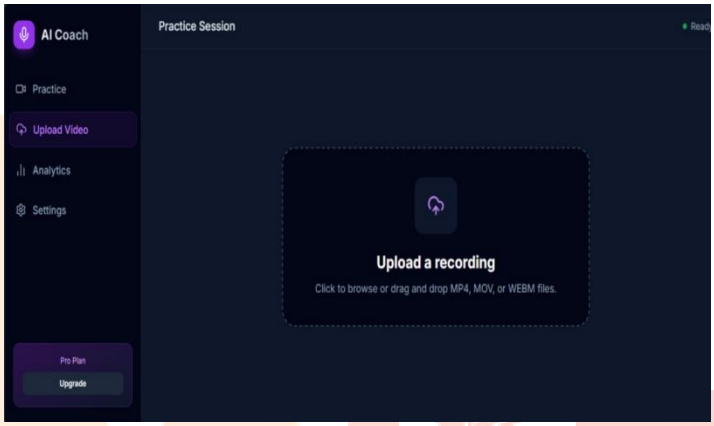


Figure 4: Upload video (Or) audio

1. Eye Contact Monitoring

MediaPipe-based landmark tracking effectively monitored gaze alignment under standard indoor lighting conditions.

Results included:

- Reliable detection when the face was fully visible
- Minor accuracy reduction in low-light settings
- Eye contact percentage metric accurately reflected engagement consistency

Moreover, users were more aware of the direction of their gaze, allowing for better visual focus.

1. Nervousness Detection (Head Movement Analysis)

Head jitter modeling based on landmark displacement variance effectively identified nervous behaviour.

Findings included:

- High movement variance accurately triggered a Red state
- Controlled head posture maintained a Calm (Green) state
- Detection remained reliable even with moderate camera resolution

This feature was found to be especially useful in mock interview simulation situations. Generally, nervous behavior is shown by the occurrence of involuntary movements. The system could differentiate between a stable position and a jitter-prone position by quantifying the repetitive movements mathematically, thus identifying confident and anxious states.

D. Integrated traffic-light feedback performance

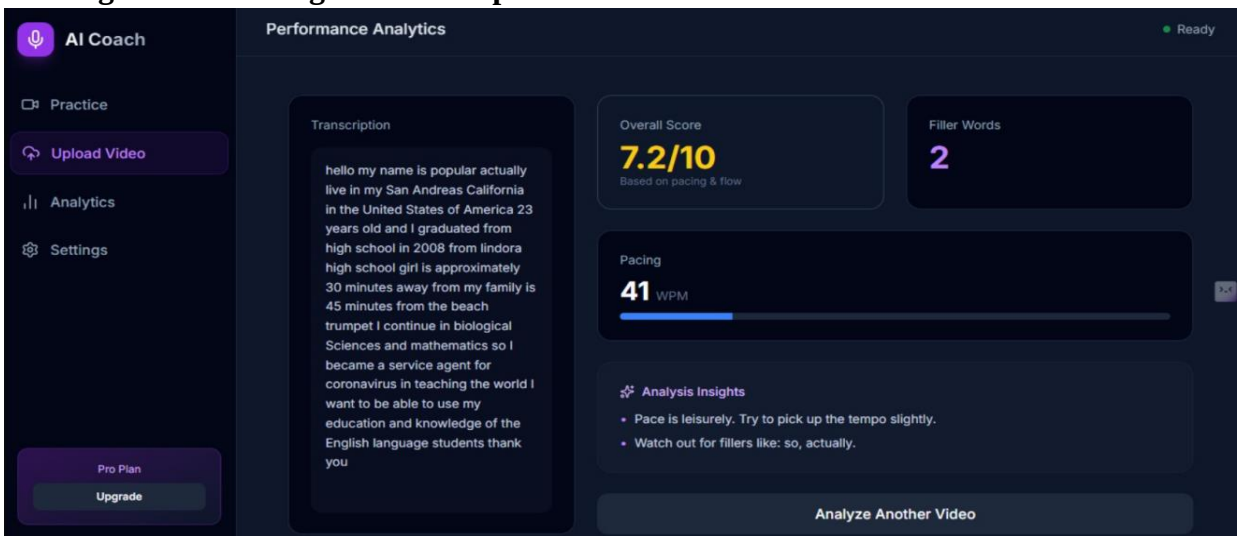


Figure 5:The real-time traffic-light system proved to be the most impactful feature.

Performance Observations:

- Green indicators reinforced stable behavior
- Yellow served as a cautionary alert
- Red delivered immediate corrective feedback

Users noted that receiving instant visual cues allowed them to adjust their behavior more quickly, offering faster self-correction compared to conventional post-session reports and improving overall engagement during live coaching sessions.

E. User feedback analysis

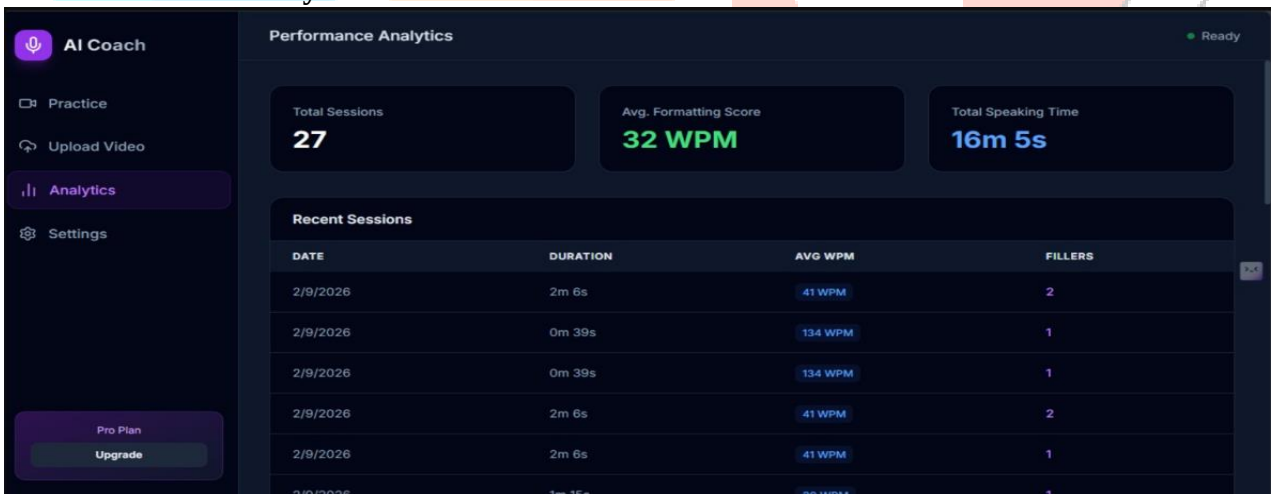


Figure 6:Feedback was collected from trial users after multiple sessions.

Key Observations:

- 85% of users noticed better awareness of their speaking pace
- 78% reduced filler word usage after multiple sessions
- 82% felt more confident sustaining eye contact
- Users valued the intuitive interface and live analytics dashboard

The integration of real-time feedback with historical analytics greatly boosted learning engagement.

F. Performance metrics summary

Parameter	Result
Real-time Latency	< 1 second
Filler Detection Accuracy	~92%
Eye Contact Detect-ion	Stable in normal lighting
System Stability	No crashes during 30+ sessions
Database Logging	Accurate session storage

G. Discussion

The Results show that combining speech analysis and computer vision within a single system greatly improves the effectiveness of communication training. The real-time traffic-light feedback sets this system apart from traditional post-session evaluators.

The system successfully achieves its main goals:

- Low-latency performance
- Accurate speech and behaviour analysis
- Scalable architecture
- User-friendly interface
- Objective and measurable evaluation

Although lighting and microphone quality may have some effect on the level of precision achieved, in regular conditions, it works effectively. In conclusion, it can be said that AI-Powered Communication Coach offers an effective and scalable communication skill improvement solution with intelligent guidance.

V. CONCLUSION

The AI-Powered Communication Coach is an model that shows the potential of Artificial Intelligence, Natural Language Processing, and Computer Vision to come together in one single system that has the potential to enhance communication skills, both verbal and non-verbal. Unlike other techniques, this system promises objective, consistent, and instant feedback in the absence of any human evaluators.

The framework can capture synchronized audio and video inputs, analyzing speech-related parameters such as Words Per Minute (WPM) and filler word counts, as well as non-verbal parameters such as eye contact and head stability. These parameters are then unified using a structured scoring system, resulting in a real-time traffic light system.

As the experimental evaluation indicates, the system offers low-latency responses (i.e., below one second), stable facial landmark detection in normal conditions, and high accuracy in filler word detection. Moreover, the analytics dashboard enables learning over time by storing the data from the session and allowing the users to monitor improvements in long-term performance.

Key contributions of this work include:

- Development of a real-time, multi-model communication analysis system
- Implementation of an interpretable, rule-based scoring model
- Integration of verbal and non-verbal assessment in a unified platform
- Introduction of a real-time traffic light feedback mechanism

- Creation of a scalable cost-effective alternative to conventional coaching.

This project has demonstrated the potential of the use of AI technology for the improvement of user awareness, confidence, and performance in the speaking skills area. It has provided a great solution for the students, the professionals, the job seekers, and the academic community at large by effectively blending technology and usability.

In conclusion, the AI-Powered Communication Coach is a great step toward the development of intelligent, automated, and accessible communication training. It is a great solution for the improvement and enhancement of the skills of the individuals interested in the effective communication.

VI. REFERENCES

[1] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L.,Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is All You Need. In Proceedings of the 31st Conference on Neural Information Processing Systems (NeurIPS).

[2] Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., & Sutskever, I. (2022). Robust Speech Recognition via Large-Scale Weak Supervision. arXiv preprint arXiv:2212.04356.

[3] Zhang, F., Bazarevsky, V., Vakunov, A., Tkachenka, A., Sung, G.,Chang, C. L., & Grundmann, M. (2020). MediaPipe Face Mesh:
Real-Time Face Landmark Detection. Google Research.

[4] Jurafsky, D., & Martin, J. H. (2021). Speech and Language Processing (3rd ed.). Pearson Education.

[5] Ekman, P., & Friesen, W. V. (1978). Facial Action Coding System:A Technique for the Measurement of Facial Movement.Consulting Psychologists Press.

[6] Rosenberg, A., & Hirschberg, J. (2009). Charisma perception from text and speech. Speech Communication.

[7] McNeill, D. (1992). Hand and Mind: What Gestures Reveal about Thought. University of Chicago Press.

[8] FastAPI Documentation. (2023). FastAPI Framework for Building APIs with Python. Available at: https://fastapi.tiangolo.com

[9] MediaPipe Documentation. (2023). MediaPipe Solutions and Face Mesh Model Description. Available at:
https://developers.google.com/mediapipe

[10] OpenCV Documentation. (2023). Open Source Computer Vision Library. Available at: https://opencv.org