IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 8.066 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 12, Issue 5, May 2025

DOI: 10.17148/IARJSET.2025.125258

Feedback Mechanism on Public Speaking using Audio and Video Analysis

Siddaraj M G[1], Abrar Khan[2], Ankith Gowda B H[3], Daivik R[4], P G Nithin[5]

Assistant Professor, Department of ISE, Maharaja Institution of Technology Mysore, Mandya, India [1]

Student, Department of ISE, Maharaja Institution of Technology Mysore, Mandya, India[2-5]

Abstract: This project introduces an innovative real-time feedback software aimed at enhancing public speaking skills through comprehensive analysis of webcam data. The system evaluates key aspects of body language such as posture, gestures, and eye contact, along with critical speech metrics including filler word usage, speaking pace, and clarity. By delivering instant, actionable feedback and detailed progress reports, it enables users to systematically improve their presentation skills. The software is built using Streamlit for a responsive user interface and backend, a Convolutional Neural Network (CNN) for analyzing non-verbal communication, Hugging Face models for advanced natural language processing, and Librosa for audio analysis and transcription. Trained on a diverse dataset of annotated public speaking videos, the system ensures high accuracy and relevance while maintaining strict privacy and ethical standards. Extensive testing has validated its reliability, and continuous updates based on user feedback allow the software to evolve with technological advancements and user needs. This AI-powered tool represents a significant step forward in making high-quality public speaking training accessible to all.

Keywords: The main keywords of the project are public speaking, real-time feedback, body language, speech analysis, CNN, Hugging Face, Librosa, NLP, audio-visual processing, feature extraction, user interface, Streamlit, Tkinter, machine learning, deep learning, emotion detection, posture, gestures, eye contact, and filler word

I.INTRODUCTION

Public speaking is a fundamental communication skill that significantly impacts personal, academic, and professional success. Whether it's delivering a classroom presentation, participating in interviews, or addressing large audiences, the ability to speak clearly and confidently is essential. Despite its importance, many individuals face difficulties in improving their public speaking due to the absence of structured, timely, and personalized feedback. Traditional feedback methods such as coaching sessions, peer reviews, and workshops are often expensive, time-consuming, subjective, and inaccessible to a larger audience, especially students and early-career professionals.

To bridge this gap, this project proposes an innovative AI-powered feedback system that delivers real-time, comprehensive analysis of public speaking performance through the integration of audio and video data. The system evaluates both verbal and non-verbal communication cues—such as speech clarity, filler word usage, speaking pace, emotional tone, posture, gestures, and eye contact—providing users with instant, actionable insights. This dual-modality approach ensures a holistic evaluation, enabling users to identify areas of improvement and track progress over time.

The system leverages a combination of cutting-edge technologies including Convolutional Neural Networks (CNNs) for body language analysis, Hugging Face models for natural language processing (NLP), and Librosa for audio analysis and speech processing. A user-friendly interface developed using Streamlit and Tkinter ensures seamless interaction, even for non-technical users. Designed with scalability, reliability, and accessibility in mind, the system also emphasizes data security and user privacy.

IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 8.066 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 12, Issue 5, May 2025

DOI: 10.17148/IARJSET.2025.125258

II.LITERATURE SURVEY

The literature survey explores a range of studies that collectively support the development of an AI-based feedback system for public speaking using audio and video analysis.

[1] introduces a feedback mechanism for customer service using Speech Emotion Recognition (SER). It employs deep learning models to analyze customer emotions from speech, using features such as pitch, tone, and intensity. The system enhances service quality by identifying emotional trends without explicit feedback, though it faces challenges such as dataset bias and high computational needs for real-time processing.

[2] focuses on detecting stuttered speech using Convolutional Neural Networks (CNNs). It processes audio signals, filters noise, extracts features using MFCC, and classifies speech fluency. Trained on the UCLASS dataset, it highlights the need for robust filtering and improved generalization due to variability in speech patterns across individuals.

[3] reviews audio-visual speech enhancement and separation, where deep learning integrates audio and video cues, such as lip movements, to improve speech clarity in noisy environments. This approach outperforms traditional audio-only models but is challenged by visual data quality, synchronization issues, and computational complexity.

[4] discusses optimizing feedback time in Voice User Interfaces (VUIs) based on users' perception. Using linear regression, the study concludes that a feedback delay of approximately 750ms improves user satisfaction. Delays beyond 1.85 seconds lead to negative emotional responses, emphasizing the importance of timely interactions in voice systems.

[5] presents feature-based eye gaze tracking techniques for human-computer interaction. It uses Fast Fourier Transform (FFT) and Z-domain methods to estimate gaze direction by analyzing pupil and iris movement. The system supports applications in virtual reality and assistive technologies but is limited by environmental conditions and user variability.

[6] proposes a multi-view approach to audio-visual speaker verification. It combines facial and voice features using mid-level and cross-modal fusion techniques. The system achieves high accuracy using the VoxCeleb1 dataset, although performance drops when only one modality (audio or video) is available, especially in noisy or visually poor conditions.

[7] explores eye movement monitoring for objectively ranking multimedia content. Eye trackers capture fixation points and gaze patterns to analyze student engagement. Although effective in identifying visual attention, the study is limited by small sample sizes and the lack of real-time capability.

[8] presents a Virtual Reality Exposure Therapy (VRET) SaaS solution for treating public speaking anxiety. Hosted on Microsoft Azure, the system enables remote therapy sessions with real-time control and biofeedback sensors. Despite providing immersive therapy experiences, the system faces challenges related to hardware limitations and internet dependency.

[9] introduces a real-time text and speech translation system using a CNN-based sequence-to-sequence model. Unlike traditional RNN models, CNNs allow faster processing and better parallelization. Trained on eight languages, the system supports audio and video translation but struggles with noise interference and limited performance in underrepresented languages.

[10] is a systematic literature review on video processing using deep learning. It categorizes techniques into CNN, RNN, and hybrid models used for action recognition, object detection, and motion tracking. The review highlights computational challenges, lack of diverse datasets, and the need for improved real-time processing in video-based applications.

[11] revisits SER in customer service, detailing a deep neural network model with feature extraction methods such as MFCC and data augmentation techniques. It assesses customer emotions over time but shares limitations with earlier studies, including cultural bias and high resource requirements.

IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 8.066 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 12, Issue 5, May 2025

DOI: 10.17148/IARJSET.2025.125258

[12] reviews various deep learning techniques for SER, comparing models like CNNs, RNNs, and DBNs. It emphasizes challenges like speech variability, real-time processing demands, and lack of annotated datasets, while advocating for multimodal input and transfer learning in future research.

[13] is a survey on sentiment analysis, which examines opinion mining from text using techniques like TF-IDF, Word2Vec, and Transformer models such as BERT. Despite advances, it struggles with context understanding, domain dependency, and multilingual complexities.

[14] provides a comprehensive review of SER systems, detailing stages such as preprocessing, feature extraction, and classification. It categorizes methods into traditional machine learning and deep learning approaches and discusses challenges like background noise, speaker variability, and data scarcity for non-English languages.

[15] proposes a deep learning system for analyzing classroom activities using both audio and video features. The system uses VGG16, OpenPose, and SincNet for feature extraction, followed by ensemble learning with CNN and LSTM models. Although effective, its accuracy is affected by environmental variations and imbalanced datasets.

[16] addresses digital audio forensics, focusing on microphone and environment classification using a hybrid CNN-LSTM model. It processes spectrogram features from the KSU-DB corpus to detect device and recording conditions. The model performs well but requires substantial computing power and struggles with generalization across diverse acoustic environments.

III.METHODOLOGY

The methodology for the project “Feedback Mechanism on Public Speaking using Audio and Video Analysis” involves a structured, multi-stage process combining machine learning techniques with audio and video processing to provide real-time, personalized feedback to users on their public speaking performance. Initially, the system begins with data acquisition, where users either upload or directly record videos of their speeches through a user-friendly graphical interface built using Tkinter. The system supports common video formats such as MP4. Once a video is submitted, the audio is extracted using the MoviePy library to enable separate and focused processing of both modalities.

The next phase involves preprocessing the collected data. For audio, preprocessing is done using the Librosa library to perform noise reduction, normalization, and segmentation. Important features such as pitch, tone, speech rate, and Mel-Frequency Cepstral Coefficients (MFCCs) are extracted. Simultaneously, the video is preprocessed by extracting individual frames, resizing them, and detecting relevant facial and body features using computer vision techniques.

In the feature extraction phase, a range of audio and visual characteristics are identified. Audio analysis includes detecting filler words using Google Speech Recognition and NLP parsing, calculating speech pace (words per minute), and analyzing emotional tone with Hugging Face NLP models. Visual analysis involves assessing body language,such as posture, hand gestures, facial expressions, and eye contact, using Convolutional Neural Networks (CNNs) and pose estion

IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 8.066 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 12, Issue 5, May 2025

DOI: 10.17148/IARJSET.2025.125258

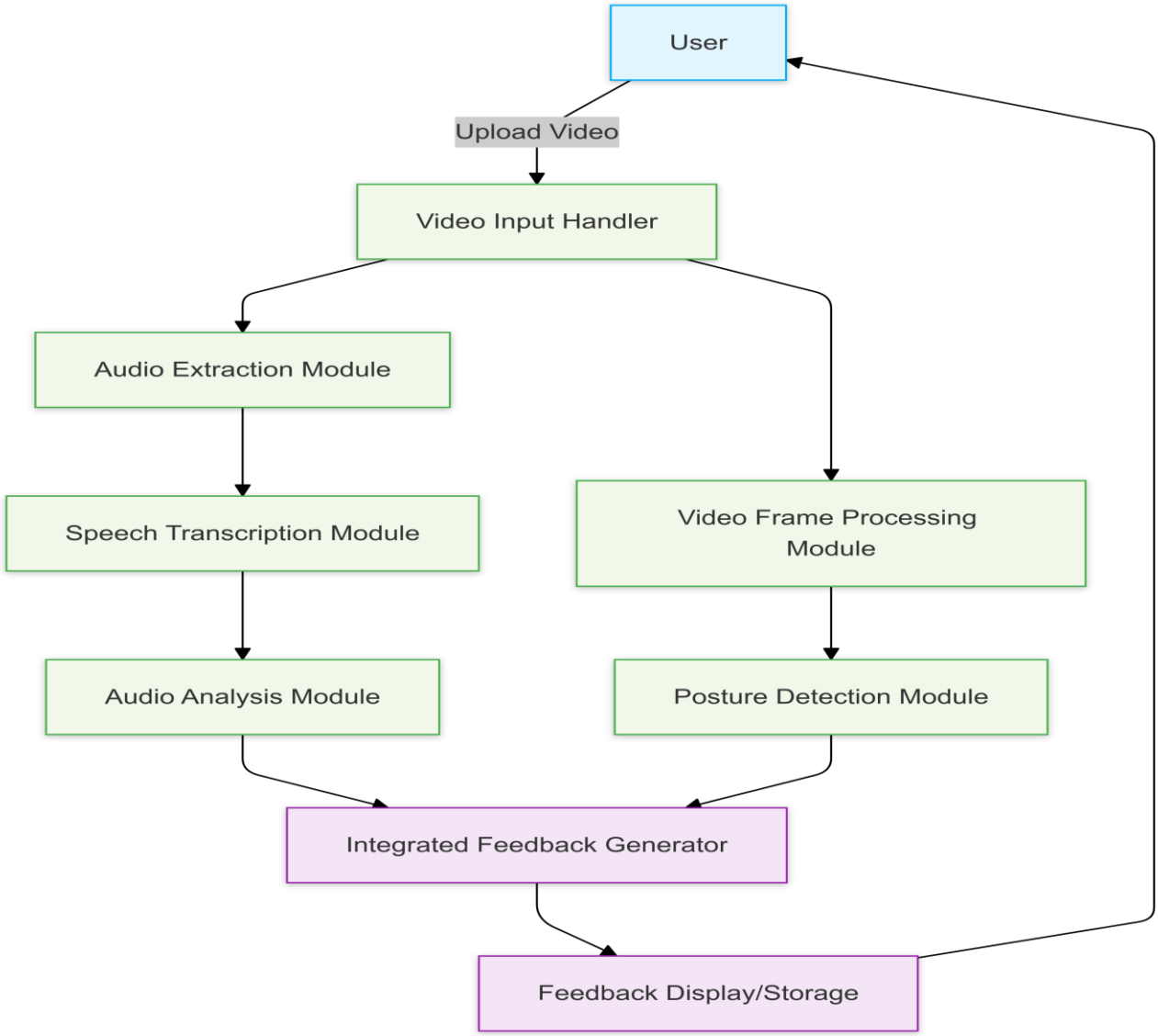

Fig 1.1 : System Design

Once features are extracted, machine learning models are applied for in-depth analysis. A CNN is used to evaluate body language and posture, classifying them as appropriate or needing improvement. Deep learning models also assess speech clarity, emotional expression, and fluency. Additionally, Error-Level Feedback Analysis (ELFA) is performed by comparing original and recompressed audio signals to detect inconsistencies in modulation, pitch, and articulation, highlighting areas of weakness in the speaker's delivery.

Following the analysis, a feedback generation module compiles all results to produce detailed, actionable insights. This includes feedback on specific aspects like eye contact, filler word usage, tone variation, and pacing. A performance score is also generated to summarize the speaker’s effectiveness. The results are then displayed through the graphical interface in both visual and textual formats, ensuring clarity and accessibility. The entire system is integrated using a Flask backend that handles API requests, ensuring modular communication between components. Designed for scalability and ease of use, the system supports real-time analysis, making it a practical tool for improving public speaking skills across various user groups.

IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 8.066 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 12, Issue 5, May 2025

DOI: 10.17148/IARJSET.2025.125258

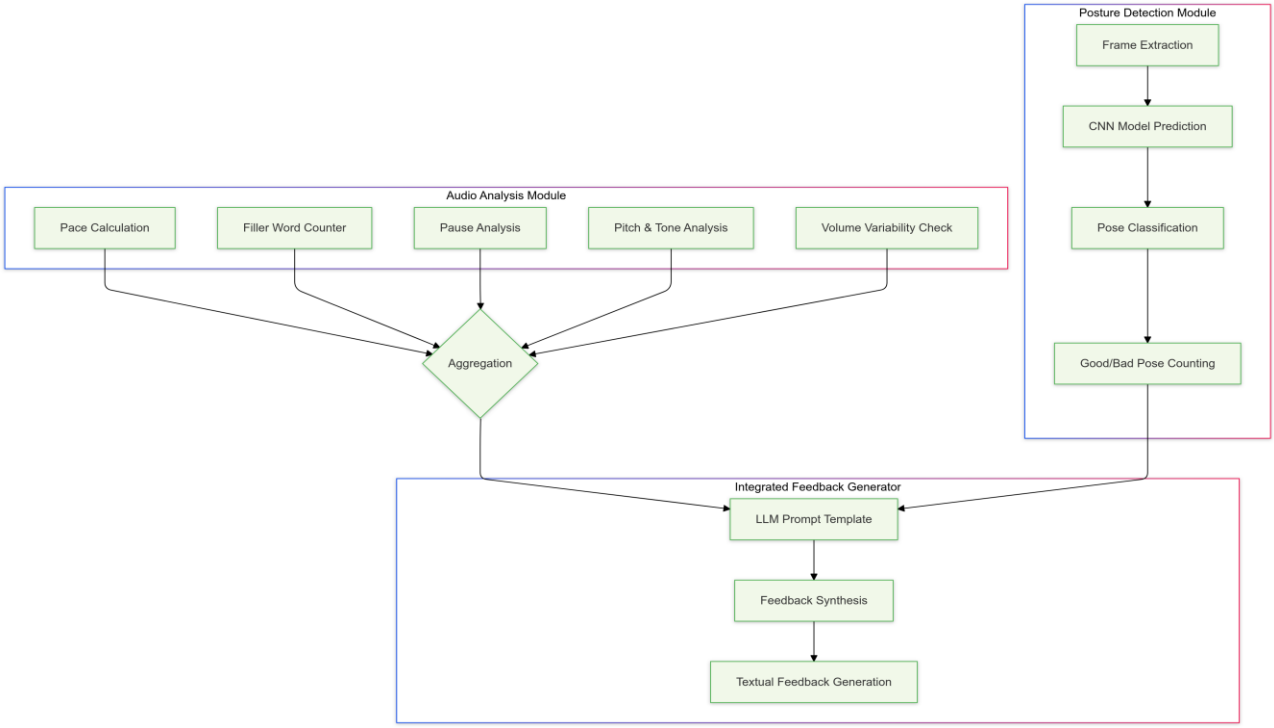

Fig 1.2 : Data flow diagram

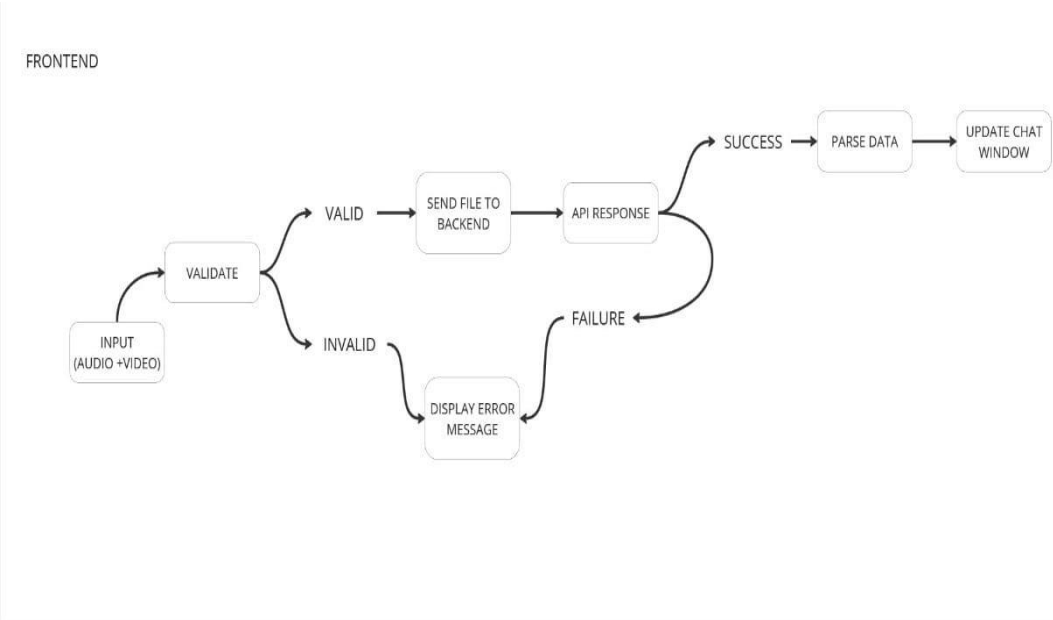

Fig 1.3 : Work flow Diagram

IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 8.066 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 12, Issue 5, May 2025

DOI: 10.17148/IARJSET.2025.125258

IV.EXISTING SYSTEM AND LIMITATION

The existing system for public speaking evaluation primarily relies on traditional methods such as in-person coaching, peer reviews, self-assessment through recorded videos, and structured workshops. While these approaches have been widely used, they come with several limitations that hinder their effectiveness and accessibility. One major drawback is the high cost associated with professional coaching and organized training programs, making them less accessible to students or individuals from economically diverse backgrounds. Additionally, the feedback provided in such systems is often subjective and inconsistent, as it largely depends on the evaluator's experience and personal biases. These methods also lack the ability to offer immediate, real-time feedback, which is crucial for speakers to make timely corrections and improvements. Furthermore, traditional systems do not typically incorporate detailed analysis of non-verbal cues such as body language, eye contact, or vocal tone, which are essential elements of effective communication. As a result, many individuals struggle to gain actionable insights into their performance, limiting their progress in developing strong public speaking skills.

Limitation of Existing System

[1] **High Cost**: Professional coaching and workshops can be expensive, making them inaccessible for many individuals, especially students.

[2] **Subjective Feedback**: Peer reviews and human evaluations are often biased and inconsistent, leading to unreliable assessments.

[3] **Lack of Real**-Time Insights: Traditional methods do not provide immediate feedback, which delays improvement and reduces effectiveness.

[4] **Limited Analysis of Non-Verbal Cues**: Body language, gestures, posture, and eye contact are rarely analyzed in depth, despite being crucial for effective communication.

[5] **Time-Consuming**: Manual assessment of presentations takes significant time and effort from both the speaker and the evaluator.

V.ALGORITHM

The algorithm implemented in the project begins by verifying whether a user has uploaded a valid video file. Once confirmed, the system extracts the audio component from the video using the MoviePy library. This audio is then processed using Librosa to extract key features such as pitch, tone, and duration. Following this, Google Speech Recognition is employed to transcribe the speech into text. The transcribed text is scanned for filler words such as "um," "uh," "like," and "you know," which are counted to evaluate the speaker’s fluency. The system also calculates the words per minute (WPM) to assess the pace of speech delivery.

To evaluate emotional tone, the system utilizes deep learning models from Hugging Face that analyze variations in pitch, intensity, and modulation, identifying patterns that suggest nervousness, confidence, or enthusiasm. This feedback is dynamically reflected using a progress bar in the interface. Based on the emotional analysis, the system indicates whether the speaker’s tone is effective or requires improvement. In parallel, the video is analyzed frame-by-frame using Convolutional Neural Networks (CNNs) to detect and classify body posture and gestures. The system identifies if the posture is upright or slouched and whether gestures such as hand movements are present or lacking.

Eye contact and body orientation are also evaluated to provide a complete view of the speaker’s non-verbal communication. An additional component, Error-Level Feedback Analysis (ELFA), compresses and compares audio to detect modulation inconsistencies and articulation clarity. These differences are amplified and visualized using waveform plots to highlight speech delivery issues. Finally, the system compiles the audio and video analysis results to generate real-time, actionable feedback. This includes specific suggestions for improvement, visual summaries, and overall performance scores, enabling users to refine their public speaking skills effectively

IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 8.066 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 12, Issue 5, May 2025

DOI: 10.17148/IARJSET.2025.125258

VI.RESULT AND DISCUSSION

The developed system successfully analyzes public speaking videos to provide real-time, personalized feedback using audio and video analysis techniques. During testing, the system was able to accurately detect filler words, assess speech pace, and identify emotional tone through audio input. It also evaluated visual aspects such as posture, gestures, and eye contact using video processing. The feedback generated was specific and actionable, highlighting areas for improvement such as excessive use of filler words, poor posture, lack of eye contact, or inconsistent speech pace. Users received visual indicators along with text-based suggestions, which helped them understand their strengths and weaknesses. In multiple test scenarios, the system showed a high level of consistency in detecting and reporting speaking flaws. Compared to traditional feedback methods like peer review or manual coaching, this AI-based approach provided faster and more objective insights. Moreover, the user-friendly interface allowed even non-technical users to interact with the system easily. Overall, the results demonstrate that the system is effective in enhancing public speaking skills, offering a scalable and accessible solution for students, professionals, and educators.

VII.CONCLUSION

In conclusion, the project successfully developed an AI-powered feedback system that evaluates public speaking performance through comprehensive audio and video analysis. By integrating machine learning techniques, the system provides real-time, personalized feedback on crucial aspects such as speech clarity, pacing, filler word usage, emotional tone, posture, gestures, and eye contact. The dual-modality approach ensures a holistic assessment of both verbal and non-verbal communication skills. The system addresses the limitations of traditional feedback methods by offering an accessible, objective, and scalable solution that can be used by individuals at any skill level. Its intuitive interface and detailed feedback reports make it a practical tool for continuous self-improvement in public speaking. The results from testing demonstrate the system's reliability and effectiveness in identifying key improvement areas, thereby helping users enhance their confidence and communication abilities. This solution holds strong potential for use in educational, professional, and training environments where effective speaking is essential.

REFERENCES

[1] IEEE-2021: Feedback Mechanism for Customer Care Service via Speech Emotion Recognition.

[2] IEEE-2023: Efficient Recognition and Classification of Stuttered Word from Speech Signal using Deep Learning Technique.

[3] IEEE-2021: An Overview of Deep-Learning-Based Audio-Visual Speech Enhancement and Separation.

[4] IEEE-2024: Optimization of Feedback Mechanism of Voice User Interfaces Based on Time Perception.

[5] IEEE-2022: Feature Based Methods for Eye Gaze Tracking.

[6] IEEE-2023: A Multi-View Approach to Audio-Visual Speaker Verification.

[7] IEEE-2021: Eye Movement Monitoring for Multimedia Content Ranking.

[8] IEEE-2024: Public Speaking Designing Software as a Service Solution for a Virtual Reality Therapy.

[9] IEEE-2024: Real-Time Text & Speech Translation Using Sequence To Sequence Approach.

[10] IEEE-2022: Video Processing Using Deep Learning Techniques: A Systematic Literature review.

[11] IEEE-2022: Feedback Mechanism for Customer Care Service via Speech Emotion Recognition.

IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 8.066 ⁜ Peer-reviewed / Refereed journal ⁜ Vol. 12, Issue x, Month 2025

DOI: 10.17148/IARJSET.2025.12xx

[12] IEEE-2022: Speech Emotion Recognition Using Deep Learning Techniques.

[13] IEEE-2024: A Survey on Sentiment Analysis.

[14] IEEE-2023: A Comprehensive Review of Speech Emotion Recognition Systems.

[15] IEEE-2022: Analysis of Classroom Processes Based on Deep Learning With Video and Audio Features.