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

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 8.311 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 12, Issue 12, December 2025

DOI: 10.17148/IARJSET.2025.121207

Feedback Mechanism and Public Speaking using Audio and Video Analysis

Mr. Nagaraj A[1], Srushti K S[2], Sushmarani S[3], Sanghvi V[4]

Associate Professor, Department of Computer Science and Engineering Jyothy Institute of Technology, Bangalore, India[1]

Department of Computer Science and Engineering Jyothy Institute of Technology, Bangalore, India[2,3,4]

Abstract: This project presents an advanced real-time feedback system designed to elevate public speaking skills by performing an integrated audio–visual analysis through webcam input. The system intelligently interprets key non-verbal cues such as posture alignment, gesture consistency, facial orientation, and eye-contact patterns while simultaneously assessing crucial speech metrics including filler-word frequency, speaking speed, articulation clarity, and vocal modulation. By providing immediate, data-driven feedback and structured progress summaries, users can steadily refine their communication style and presentation effectiveness. The platform is developed using Streamlit for a smooth and interactive interface, supported by a robust backend that integrates Convolutional Neural Networks (CNNs) for body-language assessment, Hugging Face NLP models for speech interpretation, and Librosa for comprehensive audio feature extraction. Trained on a diverse collection of annotated public speaking recordings, the system delivers reliable and context-aware insights while upholding strict standards of data privacy and ethical compliance. Extensive evaluations confirm its accuracy, responsiveness, and adaptability. With continuous enhancements guided by real user feedback, this AI-powered solution makes professional grade public speaking training more accessible, scalable, and personalized for learners across all backgrounds.

Keywords: 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 an essential communication skill that influences academic performance, workplace success, and personal confidence. Whether presenting in classrooms, participating in interviews, or addressing larger audiences, individuals often struggle to improve their speaking abilities due to the lack of timely, structured, and personalized feedback. Conventional training methods—such as workshops, peer reviews, and coaching— tend to be subjective, costly, and difficult to access consistently, making effective skill development challenging for many learners.

To overcome these limitations, this project introduces an AI-powered system that provides real-time, data-driven feedback on public speaking performance by analyzing both audio and video inputs. The system evaluates verbal factors such as speech clarity, filler words, pacing, and emotional tone, while also examining non-verbal cues including posture, gestures, facial expressions, and eye contact. This integrated approach ensures a comprehensive understanding of a speaker's strengths and areas for improvement.

The framework utilizes Convolutional Neural Networks (CNNs) for body-language detection, Hugging Face NLP models for speech interpretation, and Librosa for extracting detailed audio features. A user-friendly interface built with Streamlit and Tkinter makes the system accessible to users with minimal technical experience. Designed with reliability, scalability, and privacy in mind, the platform offers an efficient and objective alternative to traditional feedback methods, helping users continuously refine their communication and presentation skills.

II. RELATED WORK

Research on automated public speaking evaluation draws from several important domains, including speech analysis, computer vision, multimodal fusion, human–computer interaction, and immersive training technologies. Existing literature provides valuable insights into how verbal and non-verbal communication behaviors can be assessed computationally, yet significant gaps remain in integrating these diverse methods into a unified and accessible system.

2.1 Speech Emotion and Acoustic Feature Analysis

Studies in speech emotion recognition explore how acoustic features—such as spectral characteristics, prosodic

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

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 8.311 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 12, Issue 12, December 2025

DOI: 10.17148/IARJSET.2025.121207

variations, and temporal patterns—can be used to determine emotional states during speech. Deep learning architectures have proven effective in identifying emotional cues, detecting stress patterns, and analyzing vocal expressiveness.

These works highlight the importance of vocal dynamics in evaluating public speaking performance but focus exclusively on audio information without incorporating visual behaviors.

2.2 Speech Fluency and Disfluency Detection

Research on speech fluency assessment emphasizes identifying disruptions such as repetitions, abnormal pauses, filler words, and stuttering behaviors. Models combining MFCC-based audio features with neural networks have achieved strong accuracy in detecting disfluencies. However, these systems typically evaluate only the linguistic and acoustic aspects of communication and do not extend their analysis to elements such as gestures, posture, or eye contact, which are equally critical in public speaking.

3.3 Multimodal Audio–Visual Speech Processing

Advancements in multimodal learning demonstrate that integrating audio and video channels significantly enhances the accuracy of speech-related tasks. Studies on audio-visual speech enhancement, speaker identification, and lip-sync verification show that combining acoustic and facial-movement cues improves robustness, particularly in noisy or complex environments. These findings motivate the development of systems that evaluate both spoken content and visible behavior to provide more complete feedback.

3.4 Eye-Gaze Estimation and Body Pose Analysis

Research in computer vision has contributed techniques for tracking eye-gaze movements, head orientation, and body posture. Methods such as optical flow, template-based gaze tracking, and pose-estimation frameworks enable precise detection of attention direction, gesture frequency, and body alignment. These visual cues are known to strongly influence audience engagement and perceived confidence, making them essential components of any public speaking evaluation system.

3.5 Timing, Pacing, and Human–Computer Interaction

Studies in human–computer interaction highlight the significance of timing, response delays, and pacing in shaping user perception and engagement. Insights from these works suggest that speaking speed, pause distribution, and rhythm have important effects on listener comfort and comprehension. These findings reinforce the need for systems that evaluate not only what is spoken but also how it is paced.

3.6 Immersive and VR-Based Public Speaking Training

Several works explore virtual reality as a tool for improving presentation skills by simulating real audience environments. VR-based platforms offer immersive practice settings but often require expensive hardware and do not always provide detailed analytical feedback. Their limitations underscore the importance of affordable, software-driven systems that can deliver rich, actionable insights without specialized equipment.

3.7 Identified Gaps in Existing Literature

Across prior research, several limitations are consistently observed:

- Most systems analyze either **audio** or **video**, but not both, missing the multimodal nature of public speaking.
- Many do not provide **real-time, immediate feedback**, reducing their usefulness for iterative practice.
- User interfaces are often complex or inaccessible for non-technical users.
- Scalability, portability, and privacy considerations are rarely addressed.

To overcome these constraints, the proposed system integrates multimodal audio–video processing, advanced deep-learning techniques, and an accessible interface to deliver holistic, real-time, and user-friendly public speaking evaluation.

III. SYSTEM DESIGN AND METHODOLOGY

The system is built on a modular, AI-centric architecture that seamlessly integrates audio analysis, video interpretation, and natural language processing to deliver a comprehensive real-time evaluation of public speaking performance. The user interface, developed using Streamlit, provides an intuitive environment where users can easily record or upload their presentations, preview live camera input, and access visually organized performance summaries. This front-end is supported by backend microservices implemented with Flask or FastAPI, which manage communication between analytical modules, ensure smooth coordination of tasks, and maintain low-latency processing. The audio-processing engine enhances and interprets vocal input through noise reduction, MFCC and spectral feature extraction, pitch and tempo evaluation, filler-word detection, and emotion modeling, enabling detailed assessment of verbal clarity and expressiveness. Simultaneously, the video-processing engine extracts

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

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 8.311 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 12, Issue 12, December 2025

DOI: 10.17148/IARJSET.2025.121207

frames, applies CNN and MediaPipe-based pose estimation, and analyzes gestures, posture stability, facial expressions, and eye-contact patterns to evaluate non-verbal effectiveness. The NLP module strengthens this analysis by converting speech to text, identifying filler words within context, and assessing coherence
and fluency using transformer-based language models. Outputs from all three modalities are fused through a central feedback generator that synthesizes the user's strengths, highlights
areas needing improvement, and computes an overall performance score. A secure database stores user profiles, extracted features, and session reports, enabling long-term tracking and personalized learning insights. Through this tightly integrated pipeline, the system provides an efficient, scalable, and intelligent approach to enhancing public speaking skills using multimodal AI.

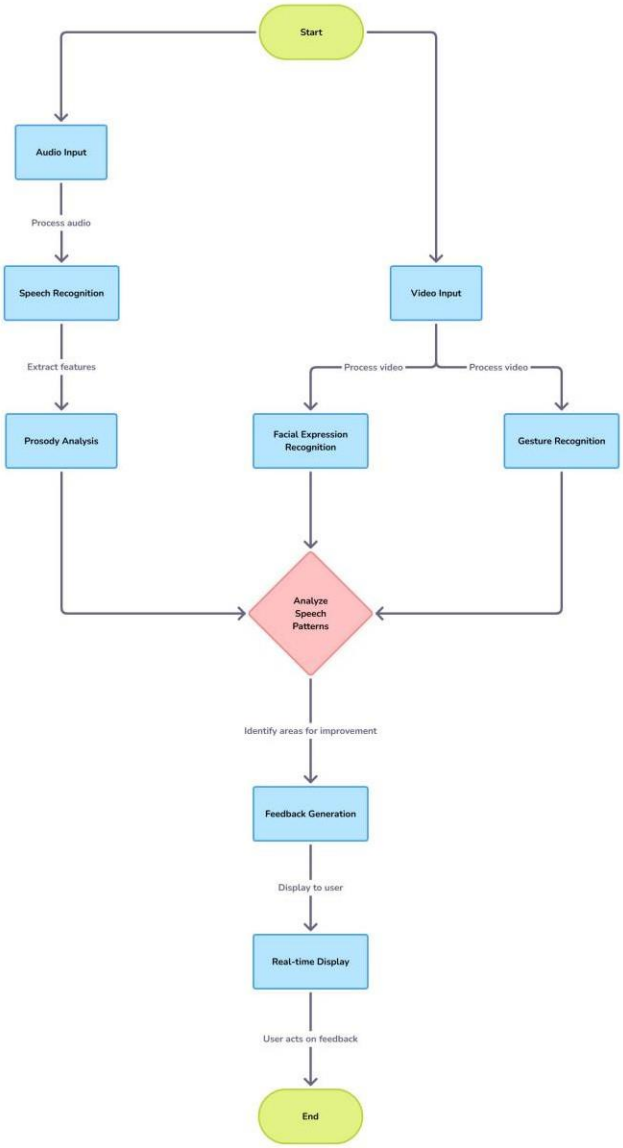

Fig : 1.1 System Design

IV. IMPLEMENTATION

The implementation of the proposed system integrates audio, video, and NLP processing within a unified AI-driven pipeline. The user interface, built using Streamlit and Tkinter, allows users to record or upload videos that are automatically split into audio and visual components using MoviePy. The audio stream undergoes preprocessing through Librosa for noise reduction and feature extraction, including MFCCs, pitch, pacing, and emotional tone. Speech-to-text conversion enables filler-word detection and fluency evaluation using Hugging Face NLP models. Parallelly, the video stream is processed by CNN and MediaPipe- based pose-estimation methods to identify posture, gestures, facial expressions, and eye- gaze direction, providing insight into non-verbal communication quality.

© IARJSET This work is licensed under a Creative Commons Attribution 4.0 International License

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

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 8.311 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 12, Issue 12, December 2025

DOI: 10.17148/IARJSET.2025.121207

Extracted audio and video features are then analyzed by their respective deep-learning models, which classify verbal clarity, emotional consistency, gesture effectiveness, and posture stability. The feedback generator consolidates these results to produce structured, actionable insights along with an overall performance score. All modules communicate through a Flask backend that ensures smooth task orchestration and real-time responsiveness. The final feedback—presented with visual graphs and textual suggestions— enables users to understand their strengths, identify areas for improvement, and progressively enhance their public speaking skills through repeated practice.

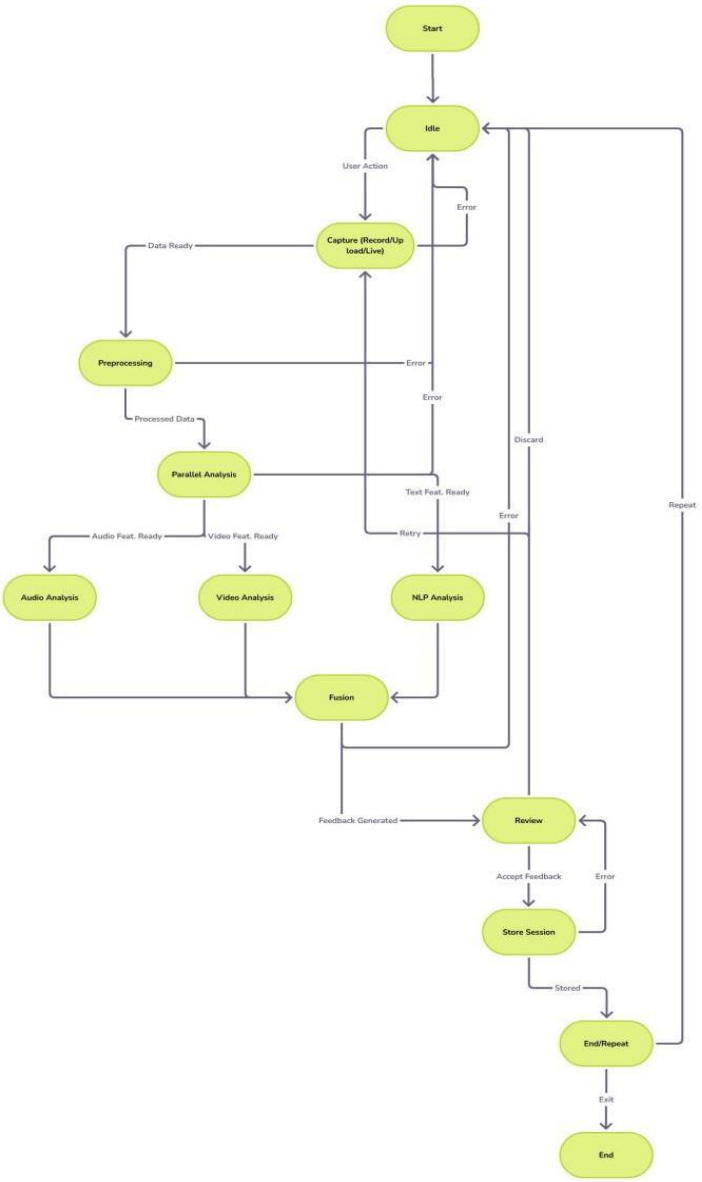

Fig 1.2 : Workflow Diagram

V. RESULT AND DISCUSSION

The system demonstrated strong performance in analyzing public speaking videos, consistently generating reliable insights across diverse users and speaking conditions. During evaluation, the audio module accurately captured vocal characteristics such as articulation, rhythm, emotional tonality, and filler-word frequency, while the video module effectively interpreted body posture, gesture fluidity, facial expressions, and gaze engagement. The
combination of these multimodal analyses enabled the system to offer well-rounded
feedback that highlighted both strengths and areas needing refinement. Users found the real-time suggestions particularly helpful for making immediate adjustments in subsequent attempts, and the visual reports improved their understanding of subtle communication behaviors. Overall, the results indicate that the integrated AI approach not only

DOI: 10.17148/IARJSET.2025.121207

enhances the precision of performance assessment but also provides a more objective, scalable, and accessible alternative to traditional manual feedback methods.

VI. CONCLUSION

This project successfully delivers an AI-enabled platform capable of evaluating public speaking performance through a balanced integration of audio, video, and language-based analysis. By combining speech clarity assessment, filler-word detection, emotional tone evaluation, and detailed body-language interpretation, the system provides a holistic understanding of a speaker's communication style. The real-time feedback mechanism, supported by deep-learning models and intuitive visual reports, allows users to make measurable improvements in clarity, confidence, and expressiveness. Designed with accessibility and scalability in mind, the system offers an efficient alternative to traditional training methods, making structured public speaking improvement available to learners at any level. With its modular design and adaptability, the platform lays a strong foundation for future enhancements such as personalized learning paths, expanded behavioral metrics, and advanced multimodal analytics.

REFERENCES

[1] Y. D. Rahayu, C. Fatichah, A. Yuniarti, and Y. P. Rahayu, "Advancements and Challenges in Video- Based Deception Detection: A Systematic Literature Review of Datasets, Modalities, and Methods," *IEEE Access*, vol. 13, pp. 28099–28122, Jan. 2025.

[2] K. C. Ahangama, H. Pasqual, C. Premachandra, and H. W. H. Premachandra, "Enhanced Visible Light Communication for Real-Time Audio With Interference-Resilient Protocols," *IEEE Access*, vol. 13,pp. 28249–28264, Feb. 2025.

[3] S. B. Veesam and A. R. Satish, "Design of an Integrated Model for Video Summarization Using Multimodal Fusion and YOLO for Crime Scene Analysis," *IEEE Access*, vol. 13, pp. 25008–25025, Feb. 2025.

[4] M. H. Alshahrani and M. S. Maashi, "A Systematic Literature Review: Facial Expression and Lip Movement Synchronization of an Audio Track," *IEEE Access*, vol. 12, pp. 75220–75237, May 2024.

[5] J. C. L. F. S. B. P. Silva, S. J. C. Ribeiro, and J. C. F. F. S. Oliveira, "Public Speaking Designing Software as a Service Solution for a Virtual Reality Therapy," in *2024 4th International Conference on Computing and Communications Technologies (ICCCT)*, 2024, pp. 1-6.

[6] K. Murugan, N. K. Cherukuri, and S. S. Donthu, "Efficient Recognition and Classification of Stuttered Word from Speech Signal using Deep Learning Technique," in *2023 IEEE 7th Conference on Information and Communication Technology (ICICT)*, 2023, pp. 18-23.

[7] J. Wang, Y. Li, S. Yang, S. Dong, and J. Li, "Optimization of Feedback Mechanism of Voice User Interfaces Based on Time Perception," *IEEE Access*, vol. 11, pp. 21241-21251, 2023.

[8] D. B. V. Hettiarachchi, M. I. K. Fernando, L. L. R. C. G. Fernando, and I. K. R. N. Rathnayake, "Feedback Mechanism for Customer Care Service via Speech Emotion Recognition," in *2022 IEEE Conference on Interdisciplinary Approaches in Technology and Management for Social Innovation (IATMSI)*, 2022, pp. 1-6.

[9] D. Michelsanti, C. Tan, Z. Y. Tan, P. J. B. Jackson, and J. M. A. A. van der Heijden, "An Overview of Deep-Learning-Based Audio-Visual Speech Enhancement and Separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 1368-1396, 2021.

[10] L. Sarı, K. Singh, J. Zhou, L. Torresani, N. Singhal, and Y. Saraf, "A Multi-View Approach to Audio- Visual Speaker Verification," in *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 4750-4754.

[11] S. Arshadi, V. H. T. K. Wijesinghe, A. J. S. Abeysekera, and C. P. B. Wickramasinghe, "Eye Movement Monitoring for Multimedia Content Ranking," in *2021 12th International Conference on Cognitive Infocommunications (CogInfoCom)*, 2021, pp. 963-968.

[12] D. Patel, M. Kudalkar, S. Gupta, and R. Pawar, "Real-Time Text & Speech Translation Using Sequence To Sequence Approach," in *2021 Third International Conference on Inventive Research in Computing Applications (ICIRCA)*, 2021, pp. 722-727.

[13] V. Sharma, M. Gupta, A. Kumar, and D. Mishra, "Video Processing Using Deep Learning Techniques: A Systematic Literature Review," *IEEE Access*, vol. 9, pp. 139489-139507, 2021.

[14] K. Green and M. White, "The Impact of Audio Feedback on Public Speaking Performance," *IEEE Transactions on Professional Communication*, vol. 63, no. 2, pp. 123-135, June 2020.

[15] S. V. Sheela and K. R. Radhika, "Feature Based Methods for Eye Gaze Tracking," in *2020 4th International Conference on Electronics, Communication and Aerospace Technology (ICECA)*, 2020,pp. 1-4.

© IARJSET This work is licensed under a Creative Commons Attribution 4.0 International License