AI-POWEREDCOMMUNICATION ENHANCEMENT TOOL ANALYSING VERBAL &NON-VERBAL CUES FOR PERSONALIZED REAL-TIME FEEDBACK

A. SRINIVASA REDDY[1], Sana Begum[2], Thathari Sowmya[3], Thallapally Abhinay [4], Ramasani Namitha[5]

[1] Assistant Professor, *Department of CSE (ARTIFICIAL INTELLIGENCE AND MACHINE LEARNING) TKR COLLEGE OF ENGINEERING & TECHNOLOGY*

[2,3,4,5] *UG Scholars in Department of CSE (ARTIFICIAL INTELLIGENCE AND MACHINE LEARNING) TKR COLLEGE OF ENGINEERING & TECHNOLOGY*

ABSTRACT: Public speaking anxiety affects countless individuals, hindering their ability to deliver messages with poise. Conventional approaches like coaching sessions or video tutorials often lack instant, tailored guidance and overlook key delivery elements such as body language, vocal delivery, and facial cues alongside content quality. This innovative AI-driven platform transforms public speaking practice by providing intelligent, user-specific feedback. It assesses speech content (via SpaCy and NLTK for natural language processing), visual performance (using OpenCV and YOLO for facial expressions and hand gestures), and vocal qualities (through OpenAI Whisper and Speech Recognition for tone, clarity, and modulation). Leveraging multimodal AI, the tool offers real-time, adaptive insights that match each user's unique style and goals. Practice sessions yield instant reports on strengths—like captivating storytelling—and improvement opportunities, such as minimizing fillers or enhancing eye contact. Designed for scalability, it handles live interactions efficiently and integrates effortlessly into web or mobile applications, making it accessible worldwide. This empowers students and professionals, particularly in underserved tier-3 areas, with data-backed coaching to overcome fears and excel confidently.

Keywords: public speaking anxiety, multimodal AI, real-time feedback, NLP, computer vision, speech analysis, personalized coaching.

1. INTRODUCTION

In schools, workplaces, and everyday life, public speaking is and a critical communication skill. Although a lot of individuals rehearse speeches, they don't get the kind of tailored feedback that accurately represents how an audience really feels. The majority of current instruments just assess fundamental speech behaviors like loudness or pace, which fail to reflect emotional reactions like if a statement is motivating, humorous, or persuasive. It is now feasible to automatically analyze speeches and comprehend audience responses thanks to developments in deep learning and the accessibility of huge speech databases online. The goal of this project is to develop a system that can analyze speech and apply emotional tags to it through affective audio annotation. The system uses cutting-edge neural networks to identify patterns in speech audio and forecast several emotions at once, enabling speakers to enhance the emotional effect of their delivery.

1.1 MOTIVATION

Although public speaking is a crucial skill, many individuals find it difficult to advance

because they don't receive constructive criticism. Traditional practice methods, such as reading novels or rehearsing in front of a mirror, do not accurately reflect how the audience is feeling. Thousands of actual speeches are now available for learning thanks to the development of AI and online videos. This supports the notion that machine learning may be used to evaluate speeches automatically and provide useful input. By comprehending how their voice affects the audience, such a system can help speakers enhance their style, self-assurance, and impact.

1.2 PROBLEM DEFINITION

The challenge is to develop an automated system that can accurately analyze the audience's emotional reactions to a public address. The system must do multi-label classification because each speech might evoke a variety of emotions, such as humor, education, or inspiration. The difficulty is that audio signals have intricate patterns, and audience emotions cannot be conveyed by low-level characteristics alone. The system has to acquire helpful intermediate-level characteristics that link speech features to audience responses. The ability to speak well in public depends on both the content of the speech and the way it is presented. Numerous current tools, like Yoodli, exclusively concentrate on vocal feedback, such as speaking rate, fluency, and filler word usage. Although this is useful, it doesn't give you a full picture of how well someone is doing. A significant role in how an audience reacts is played by crucial nonverbal indicators like facial expressions, eye contact, hand movements, posture, and body language. Because existing systems don't analyze these visual signs, speakers only get partial feedback, which hinders their progress. The primary issue is the absence of an AI system that can integrate verbal (audio/text) and nonverbal (visual/body language) characteristics to provide holistic, multi-modal feedback. Understanding audience impact, confidence, and emotions requires more than just conventional NLP and speech-processing pipelines. As a result, we need a cutting-edge deep learning-based system that can evaluate the quality of speech audio and visual motions to forecast emotional reactions and overall delivery quality.

1.3 PROPOSED SYSTEM

The suggested method is an AI-powered instrument that offers comprehensive and automated input to help users enhance their presentation abilities. The system will thoroughly analyze the user's performance, allowing them to either upload an old video or capture a fresh one. This system analyzes both verbal and nonverbal cues to provide a comprehensive assessment of the language, unlike today's tools that mostly concentrate on spoken words or speech. The system employs sophisticated Natural Language Processing (NLP) technologies like SpaCy and NLTK for verbal analysis to assess clarity, speed, filler words, and the overall structure of the content. It also measures tone, volume, pitch, and vocal variation using AI speech models such Whisper. The system uses computer vision methods employing tools like OpenCV and YOLO for non-verbal analysis. These models analyze facial expressions, body language, and hand gestures to gauge confidence, engagement, and emotional expression. Lastly, the system offers individualized feedback with performance summaries, visual highlights, and explicit recommendations. As a comprehensive solution for developing excellent public speaking abilities, this enables users to

continually enhance both the content of their speech and their delivery.

2. Usecase Diagram

The UML diagram known as a Use Case Diagram gives a general picture of the system's functional relationships with its users and external systems. It catches the functions and responsibilities of each player, as well as the application cases (functions) that define what the system should do. Use case diagrams can help define system limits, clarify functional requirements, and enhance communication between developers, analysts, and stakeholders by visualizing the system obligations and engagements.

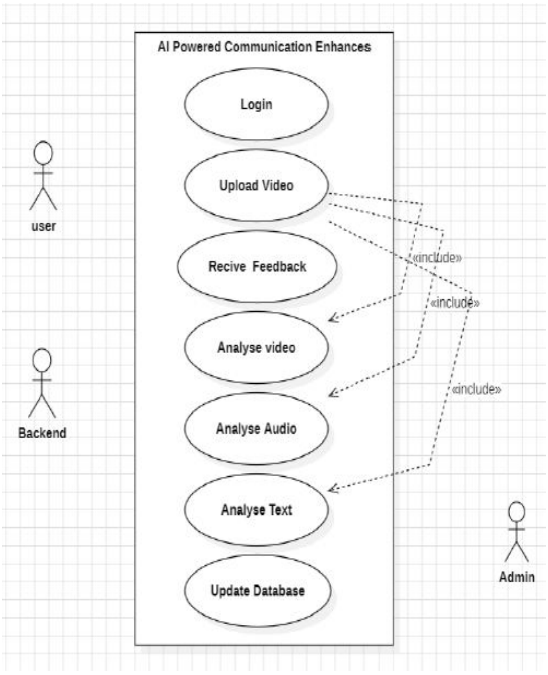

Fig 4.1.4 Use case Diagram

Sequence Diagram

A sequence diagram shows how different parts of the system interact over time. It begins when the user logs in or registers and uploads a video. The system processes the video and sends it to three modules: audio, text, and video analysis. Each module analyses its part and returns results. These results are combined to generate a final feedback report. the report is shown to the user, and the data is stored in the database.

key points

The user starts the process by uploading a video. The system prepares the data and sends it to different analysers. Each analyser processes its input and returns results. The system merges all outputs into one feedback report and displays it to the user. Finally, the results are saved for future reference.

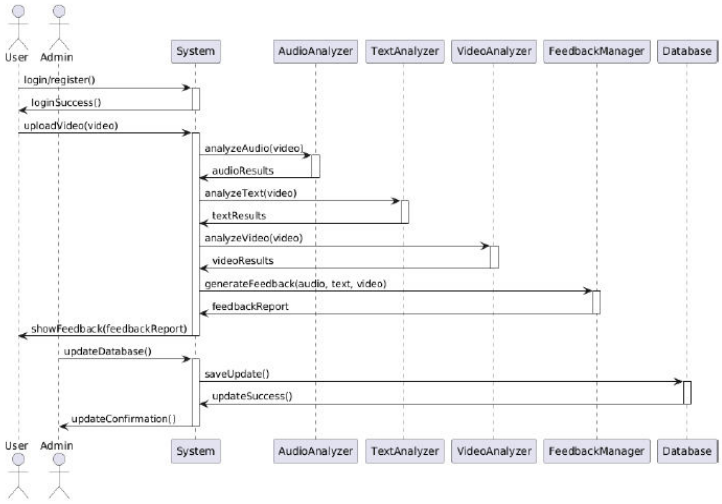

Fig.2: Sequence Diagram

3. Algorithms

The proposed system follows a multimodal AI-based approach to analyses and enhance public speaking skills. The algorithm processes the user's video input through two parallel pipelines verbal analysis and non-verbal analysis and finally fuses both results to generate an overall confidence score. The complete workflow is divided into three major phases: data preprocessing, feature extraction & analysis, and feedback generation.

System Architecture

The architecture consists of the following key modules:

- Input Module: Accepts video input (recorded or uploaded) from the user.
- Preprocessing Module: Extracts audio from the video and prepares frames for analysis.
- Verbal Analysis Module: Performs speech-to-text conversion, filler word detection, and pace calculation.
- Non-Verbal Analysis Module: Detects eye contact, posture, and hand movements using computer vision.
- Fusion & Scoring Module: Combines verbal and non-verbal scores to compute the final confidence score.
- Feedback Generation Module: Produces personalized, actionable feedback for the user.

4. IMPLEMENTATION & RESULTS

4.1 Explanation and Key Function

The core features that have been implemented in both the front and back ends of the ConfidentSpeak program are discussed in detail in this chapter.

The web-based AI software known as the ConfidentSpeak system is intended to help students develop their public speaking abilities. The verbal and nonverbal analysis pipelines process a video input, whether it be live from a webcam or uploaded from the device, and produce helpful insights as well as an overall confidence score.

The system has a client-server architecture:

The user interface, video recording, file upload, and result presentation are all managed by the front-end (React.js).

Backend (Python + FastAPI): Conducts all AI computations, such as computer vision analysis and speech transcription.

The full procedure is carried out in a modular, step-by-step manner to ensure precise analysis and a smooth user experience.

4.2 Method of Implementation

These are the main tasks performed by the application:

1. Video Input Management Function

This capability allows you to either upload a pre-existing video file or record a video using the device's webcam. The video stream is gathered and sent to the backend for analysis. Before analysis, the function guarantees that the video format is correctly verified and temporarily stored.

2. The ability to extract audio

Upon receiving the video, this method extracts the audio track from it. A 16 kHz sampling rate is used to transform the retrieved audio into a single-channel (mono) format, making it ideal for speech recognition systems. A precise transcription depends on this procedure.

3. Verbal Analysis Abilities

This is one of the system's fundamental tasks. The OpenAI Whisper small model is used for transcription. The function performs two subtasks after transcription:

Includes a list of popular filler words, such as um, uh, like, so, and many more.

determines the rate of speaking in words per minute (WPM).

The user's fluency and clarity of speech are assessed quantitatively by this function.

4. Instrument for nonverbal analysis

The visual elements of the speaker's presentation are analyzed using the MediaPipe library in this manner. In order to ascertain, it analyzes every frame of the film.

- Eye Contact: By analyzing the z-coordinates of the eye landmark, one may ascertain how long the speaker maintains eye contact with the camera.
- Posture: We can tell if the speaker is slouching or standing tall by looking at the alignment of their shoulders.
- Fidgeting: Checks for signs of worry by keeping a close eye on any unexpected motions of the hand or wrist. Knowing body language, which is a key component of successful public speaking, is facilitated by this approach.

5. The function that calculates the overall confidence score

In this function, a weighted formula combines the findings of nonverbal and linguistic analysis:

The overall confidence score is calculated using the following formula: (0. 52 verbal score) + (0. 48 nonverbal score).

The verbal score punishes the overuse of filler terms and promotes an ideal rate of speaking. The nonverbal score places greater emphasis on eye contact and demeanor, while fidgeting is penalized. The last score is limited to between 48 and 96 in order to guarantee that the values are accurate.

6. System for Producing Feedback

This method provides distinct, straightforward feedback, as shown by the calculations. It emphasizes both the performance's strengths and areas for improvement (such as reducing fillers and improving eye contact).

4.2.1 Forms

The program has an easy-to-use and straightforward video input interface. The main input form, which is located in the Practice Speaking area, includes the following elements:

Live Recording Form

The device has a big, round Start Recording button that activates its camera and microphone.

As soon as the recording starts, the button changes to Stop Recording.

In a live video preview window, the user's camera stream is shown in real time.

The recording is immediately saved as a. webm file when the user clicks. Stop.

Video Upload Form

A conspicuous Upload Video button allows users to select any previously recorded video file from their device.

Supported file formats include. webm and. mp4.

In the preview window, the uploaded video is immediately visible before analysis.

The two formats are merged into a single, straightforward user interface. After the video is prepared (either recorded or uploaded), an Analyze with AI button is displayed, which sends the video to the backend for analysis. Only legitimate video files are accepted, along with the necessary verification.

Output Screens

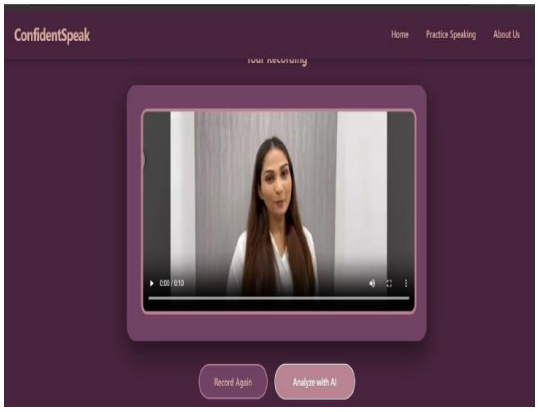

Fig 4.1 Uploading Recorded video

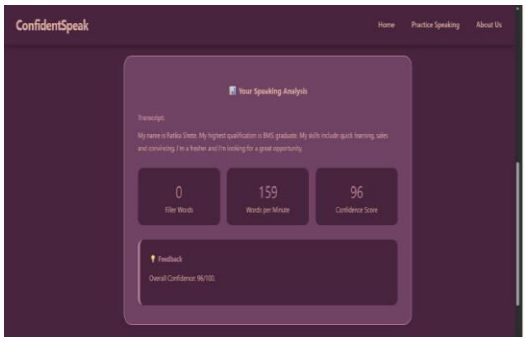

Fig 4.2 Result

Result Analysis

Once the user uploads their video, the system automatically processes it and presents the results in a clear and easy-to-read format. The evaluation is divided into two main aspects of communication.

Verbal analysis

This focuses on the way the user speaks. It includes converting speech into text, identifying the frequency of filler words, and analyzing the overall pace and clarity of speech delivery.

Nonverbal analysis

This part evaluates the user's body language and visual presentation. It checks eye contact, posture, and any unnecessary movements such as fidgeting that may affect the overall impression.

After both analyses are completed, the system calculates an Overall Confidence Score. This score is obtained by combining verbal and non-verbal performance using a weighted method, providing a quick overview of the user's confidence level.

In addition to the score, the system provides the following:

• A complete transcript of the spoken content for self-review

• Practical suggestions to improve specific areas

• Quantitative metrics that can be tracked across multiple sessions

This detailed evaluation helps users recognize their strengths and identify areas that need improvement. With regular practice and continuous feedback, users can gradually enhance their communication and public speaking skills.

5. CONCLUSION

This project introduces a straightforward way to enhance communication skills by combining technology with real-time performance analysis. Rather than focusing only on theory, it encourages users to

practice actively and receive immediate feedback on their performance. By evaluating both verbal and non-verbal elements, the system offers a more complete understanding of how effectively a person communicates.

A major outcome of this work is that it increases self-awareness in users regarding their speaking patterns, body language, and overall confidence level. Features such as speech transcription, personalized feedback, and performance scoring make the learning experience more interactive and self-guided. Users can clearly identify their mistakes and also monitor their gradual improvement over multiple practice sessions.

This system is particularly helpful for students and individuals who want to build confidence in public speaking in a stress-free setting. It provides a comfortable practice space where users can repeat sessions as needed and improve at their own pace, making the learning process more consistent and practical.

In future enhancements, the system can be extended with advanced capabilities such as emotion recognition, adaptive learning modules, and multilingual support. These additions can improve accuracy and make the platform more inclusive and effective for a wider audience.

Overall, this project demonstrates how technology can support the development of strong communication skills, helping individuals become more confident, expressive, and prepared for real-world interactions.

6. REFERENCES

[1] Xu J, Zhang B, Wang Z, Wang Y, Chen F. Affective Audio Annotation of Public Speeches with Convolutional Clustering Neural Network. IEEE Transactions on Affective Computing. 2022 Jan-Mar;13(1):238-249.

[2] Wang, T., & Mei, Y. "AI-powered Conversation Bots Enhance L2 Speaking Skills and Reduce Anxiety," Journal of Educational Technology & Society, vol. 28, pp. 123-132, 2025.

[3] Improve Communication Skills Using AI," Proceedings of the ICML Workshop on AI-assisted Learning, 2023.IEEE

[4] Pepino, M., Baskar, M. K., & Vincent, E. "Evaluating Self-Supervised Speech Representations for Speech Emotion Recognition," INTERSPEECH, 2022

[5] Zeng, Z., Pantic, M., Roisman, G. I., & Huang, T. S. "Real-Time Recognition of Affective States from Nonverbal Features of Speech and Its Application for Public Speaking Skill Analysis," Proceedings of the International Conference on Multimodal Interfaces, 2011.IEEE

[6] Li, Y., Zhang, J., & Liu, Y. "Cost-Sensitive Multi-Label Learning for Audio Tag Annotation and Retrieval," IEEE Transactions on Audio, Speech, and Language Processing, vol. 19, no. 7, pp. 2037-2048, 2011.

[7] Chollet, G., Massaro, D. W., & Graham, J. "Towards a Multimodal Virtual Audience Platform for Public Speaking Training," Proceedings of the ICMI, 2013.

[8] Hoque, M. E., Courgeon, M., Martin, J. C., Mutlu, B., & Picard, R. W. "Rich Nonverbal Sensing Technology for Automated Social Skills Training," IEEE Transactions on Affective Computing, vol. 5, no. 2, pp. 212-224, 2014.

[9] Wei, Y., Xia, W., Huang, J., Wang, Y., & Huang, T. "HCP: A Flexible CNN Framework for Multi-Label Image Classification,"

IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 38, no. 9, pp. 1901-1907, 2016

[10] Harris, A., & Hoque, M. E. "Automatic Assessment and Analysis of Public Speaking Anxiety: A Virtual Audience

Case Study," IEEE Transactions on Affective Computing, vol. 11, no. 2, pp. 222-234, 2020.

[11] Narayanan, A., & Wang, D. "Speech Emotion Recognition Using Deep Learning Techniques: A Review," Information Processing & Management, vol. 59, no. 1, 2022.

[12] Lee, S., & Lee, H. "AI-Assisted Enhancement of Student Presentation Skills," Proceedings of the International Conference on AI in Education, 2023. Smith, R., & Kaur, P. Gupta, S., & Singh,

Information Processing & Management, vol. 59, no. 1, 2022.

[13] Fernandez, Koelstra, S., & Valstar M. "Explainable AI for Audio and Visual Affective Computing: A Scoping

Review," ACM Computing Surveys, 2025.

[14] Cakir, E., Parascandolo, G., Heittola, T., & Virtanen, T. "Convolutional Recurrent Neural Networks for Polyphonic Sound Event

Detection," IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 25, no. 6, pp. 1291-1303, 2017.

[15] Cheng, M. M., Zhang, Z., Lin, W. Y., & Torr, P. H. S. "BING: Binarized Normed Gradients for Objectness Estimation at 300fps," IEEE Conference on CVPR, pp. 1531–1538, 2019.