

AI-Powered Public Speaking Coach

Conceptual Design Report

Azmi Ahmed, Elaha Ahmed, Rashmee Gade, Prachi Patel, Nethra Sakthivel

Advisor: Demetrios Lambropoulos

Team Number: SP26-45

Abstract—Many individuals struggle with public speaking due to anxiety, lack of feedback, and limited access to resources that can help improve their skills. This project aims to develop an AI-powered public speaking coach that helps users improve their confidence, clarity, and delivery. Using an iOS application, the system captures live audio and video as the user practices a speech. Machine learning models analyze key aspects of communication, including speaking pace, clarity, filler words, posture, gestures, eye contact, and facial expressions. By leveraging recent advances in multimodal AI and speech analysis, the app provides users with real-time on-screen feedback to guide them toward stronger communication habits. After each session, the system generates a summary report and tracks performance over time to highlight progress. By integrating these multimodal analyses, this project develops an accessible tool that enables individuals to become more effective speakers in academic, professional, and everyday settings.

Index Terms—Multimodal AI, Mobile Application Development, Computer Vision, Speech Processing, Natural Language Processing

I. RECOGNIZE THE NEED

Effective public speaking is a critical skill in academic, professional, and social settings, yet many individuals lack access to consistent, objective, and personalized feedback while practicing. Traditional methods for improving public speaking—such as classroom presentations, coaching sessions, or peer feedback—are often limited by availability, cost, time constraints, and subjectivity. Because of these restrictions, speakers frequently practice alone and don't get the opportunity to receive meaningful insight into their delivery, body language, tone, and content effectiveness.

Existing public speaking tools primarily focus on a single aspect such as speech pace or filler word usage, and fail to provide a holistic evaluation of a speaker's overall presentation. Additionally, most feedback is delayed or generic, making it difficult for users to immediately identify and correct weaknesses during practice sessions. Human coaching, while effective, is not always accessible or scalable for frequent practice.

There is a clear need for an affordable, accessible system that can evaluate multiple aspects of a presentation and provide actionable feedback in real time, much like a human would. An AI-powered public speaking coach addresses the gap by using computer vision, audio analysis, and natural language processing to objectively assess body language, facial expressions, vocal delivery, and content relevance. By delivering structured, data driven feedback, this system enables users to

practice independently, track improvement over time, and build confidence through informed repetition. This project responds to that need by proposing a mobile application that offers comprehensive, personalized feedback, empowering users to improve their public speaking skills anytime and anywhere.

II. DEFINE THE SCOPE OF THE PROJECT

The scope of this project is focused on designing and implementing an AI-powered mobile application that provides users with real-time and post-session feedback to improve public speaking skills. The application is intended as a practice and coaching tool rather than a replacement for professional training. It supports users as they rehearse speeches independently by offering structured, objective insights into their performance.

The system evaluates four primary dimensions of public speaking: body language, facial expressions, speech delivery, and content relevance. Using the device's built-in camera and microphone, the application captures audio and video during a practice session. This data is analyzed using multimodal AI techniques to assess posture, gestures, eye contact, facial emotion, speaking pace, pitch variation, filler word usage, and overall vocal clarity. These features were selected because they are widely recognized as key contributors to effective and engaging presentations.

Real-time feedback is limited to high-level, non-intrusive cues such as pacing alerts or posture reminders to avoid overwhelming the user during delivery. More detailed analysis is presented after the session in the form of a summary report, which includes performance scores, identified strengths, areas for improvement, and short, actionable suggestions. The application also tracks performance trends over time, allowing users to measure improvement and build confidence through repeated practice.

Content analysis is included within the scope to ensure that users remain focused on their intended message. By comparing the transcribed speech to a user-defined topic or prepared outline, the system identifies topic drift and missed points. This feature is designed to support clarity and organization rather than evaluate the correctness or quality of the speech content itself.

The scope of this project does not include live audience interaction, multilingual speech analysis, or advanced emotional coaching beyond basic confidence and engagement indicators. All analysis is designed to operate within a controlled practice

environment, prioritizing accessibility, privacy, and ease of use. The application relies on open-source tools and existing speech-to-text services and does not require specialized hardware beyond a standard smartphone.

From a system perspective, the primary inputs include live audio and video captured through the mobile device’s microphone and camera, as well as optional user-provided topic outlines. System outputs consist of real-time visual feedback cues, post-session performance summaries, and progress tracking metrics displayed within the application. Major subsystems include the frontend mobile interface, backend orchestration services, and independent machine learning modules for vision, audio, and language analysis. External services such as cloud-based speech-to-text APIs support transcription and content evaluation.

Overall, this project delivers a comprehensive yet practical public speaking coaching tool that integrates visual, vocal, and language-based feedback into a single, user-friendly mobile application. The scope is intentionally defined to ensure technical feasibility within the project timeline while still providing meaningful value to users seeking to improve their public speaking skills.

III. PRELIMINARY DESIGN

The preliminary design of this project focuses on developing a mobile application that helps users improve their public speaking presentations by providing real-time and post-session feedback. The application assesses the core aspects of public speaking, including body language and facial expressions, speech delivery, content relevance, and overall effectiveness. By combining computer vision, audio analysis, and natural language processing, the system provides immediate feedback that users can apply during practice sessions to enhance their overall skills.

The application will capture video and audio using the device’s built-in camera and microphone. This data will be processed using a combination of open-source and cloud-based tools to evaluate the user’s performance throughout a speaking session. The user interface will be developed using React Native to support both iOS and Android platforms. React Native’s permissions library will be used to manage secure access to the camera and microphone. Figure 1 presents a high-level system block diagram illustrating how user inputs are processed through the application and backend services to generate feedback and performance summaries.

To analyze body language, we will be implementing MM-Pose, an open-source pose estimation tool built on PyTorch. Video frames will be analyzed to track posture, hand gestures, head orientation, and overall movement. These features will be used to identify behaviors associated with confidence and engagement, such as controlled gestures and upright posture, as well as behaviors that may indicate nervousness or distraction. The extracted data will be translated into clear, user-friendly feedback during and after the session.

Facial expressions will be analyzed to provide additional insight into the speaker’s emotional state. DeepFace is an

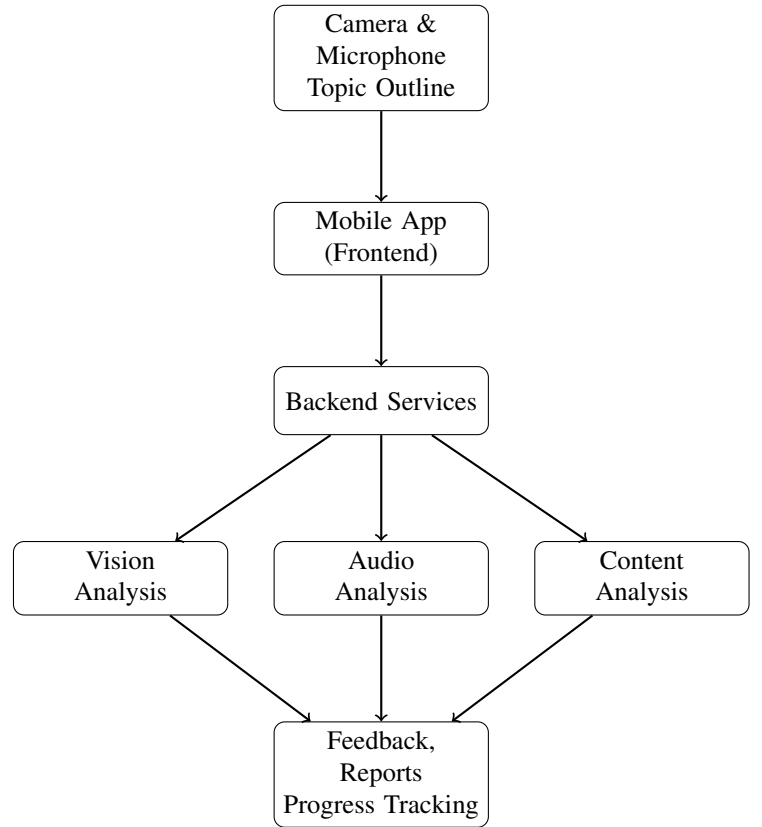


Fig. 1. System block diagram of the AI-powered public speaking coach.

open-source Python library used to detect facial emotions that indicate engagement, confidence, or anxiety. When combined with posture and movement data, facial analysis enables the system to form a more comprehensive assessment of how the speaker is perceived by the audience.

Speech delivery will be evaluated using Librosa, a Python-based audio processing library. The system will analyze pitch variation, speaking pace, pauses, and the use of filler words. These features will be used to score speech clarity and vocal confidence. Based on the results provided at the end, users will receive feedback designed to improve pacing, articulation, and vocal variation.

To evaluate speech content, the application will include speech-to-text functionality using Google Speech-to-Text. The generated transcription allows users to review their speech and serves as input for content analysis. Sentence-BERT will be used to measure content relevance by comparing the spoken text to the user’s intended topic or presentation goal. To analyze this, the user will submit their speech into the system, and the system will compare it to the two. This analysis helps identify topic drift and assess whether the speaker remains focused throughout the presentation.

The backend of the application will be implemented using Node.js and Python. Node.js will manage API requests and coordinate communication between the frontend and the machine learning modules, while Python will host the analysis

libraries. These components will communicate through REST APIs, enabling the system to combine feedback from multiple analysis modules into a unified response for the user at the end of the session. Figure 2 illustrates the end-to-end data flow of a practice session, from speech capture through multimodal analysis to feedback delivery and progress tracking.

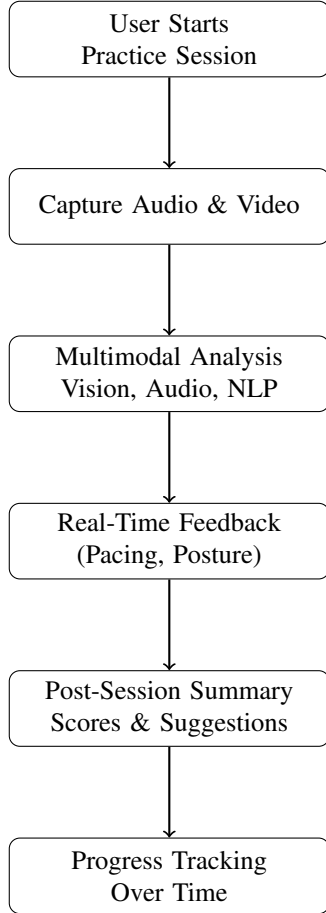


Fig. 2. Data flow of the public speaking practice and feedback process.

Overall, this preliminary design supports the development of an interactive and adaptive public speaking practice tool. By integrating visual, vocal, and content-based feedback into a single application, the system enables users to actively refine their public speaking skills and track their improvement over time, ultimately delivering a polished, well-rehearsed speech.

IV. PLAN THE PROJECT

This project follows a clear, milestone-based plan that fits the capstone timeline and supports steady development, integration, and testing. Work is divided into phases with specific goals and deliverables.

A. PHASE 1: Requirements and System Design (December)

We finalize system requirements based on the project scope, focusing on real-time feedback, post-session analysis, and progress tracking. During this phase, we design the backend

architecture, data flow between modules, and create early UI/UX wireframes to keep the app simple and easy to use.

B. PHASE 2: Module Development and Testing (January):

Each analysis module is developed and tested independently. This includes body language (MMPose), facial expressions (DeepFace), speech and audio analysis (Librosa), and speech-to-text with content monitoring (Google Speech-to-Text and Sentence-BERT). Unit testing ensures each module performs reliably before integration.

C. PHASE 3: Multimodal Integration (February):

All modules are combined into a single multimodal pipeline. The backend coordinates audio, video, and text processing in parallel, merging results into a unified feedback system using REST APIs between Node.js and Python services.

D. PHASE 4: User Testing and Refinement (March):

The integrated system is tested with real users in controlled practice sessions. User feedback and performance data are used to refine feedback logic, scoring, and the overall user interface.

E. PHASE 5: Final Evaluation and Presentation (April):

The system is evaluated against project goals, final documentation is completed, and results are presented in the final capstone report and presentation.

F. Team Responsibilities

Team responsibilities are distributed based on individual technical focus areas. Prachi Patel is responsible for speech and audio analysis, Nethra Sakthivel handles body language analysis, Rashmee Gade leads the frontend user interface development, Elaha Ahmed focuses on facial expression analysis, and Azmi Ahmed manages content monitoring and speech-to-text functionality.

G. Testing and Validation

System testing will include unit testing of individual analysis modules, integration testing across multimodal pipelines, and user-based validation to assess feedback accuracy and usability. Performance metrics such as speech rate accuracy, filler word detection reliability, and system responsiveness will be used to evaluate system effectiveness.

H. Overall

This plan ensures a balanced workload, smooth integration, and timely completion while delivering a practical and effective AI-powered public speaking coach.

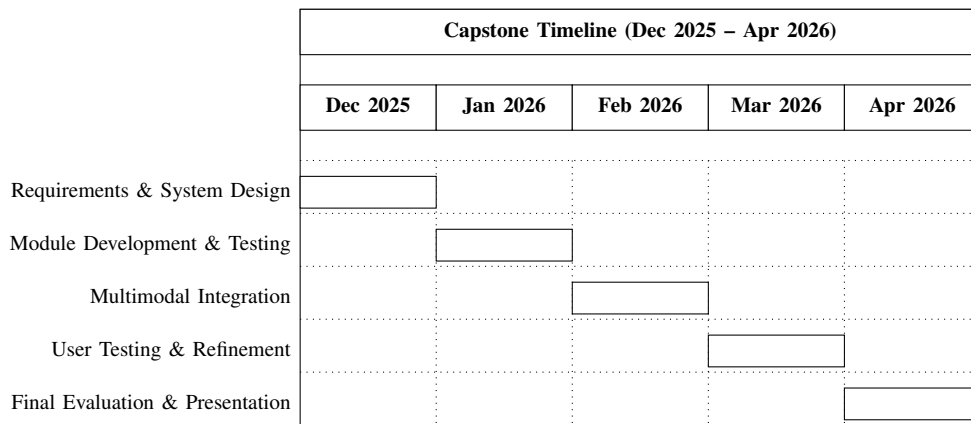


Fig. 3. Gantt chart illustrating the capstone project timeline from December through April.

V. PROJECT TIMELINE

VI. ACKNOWLEDGEMENTS

The team would like to express sincere gratitude to our faculty advisor, Prof. Demetrios Lambropoulos, for his guidance, feedback, and continuous support throughout the development of this project. His insights and encouragement played a key role in shaping the technical direction and overall scope of our work. We would also like to thank the Department of Electrical and Computer Engineering at Rutgers University for providing the resources and academic environment necessary to complete this capstone project. Finally, we thank our peers and test users for their valuable feedback, which helped refine the system and improve its usability.

REFERENCES

- [1] M. I. Tanveer, R. Zhao, and M. Hoque, "Automatic identification of non-meaningful body-movements and what it reveals about humans," arXiv:1707.04790, 2017.
- [2] A.-T. Nguyen, W. Chen, and M. Rauterberg, "Online feedback system for public speakers," in Proc. IEEE IS3e, 2012, pp. 1–5.
- [3] T. Michelson and S. Peleg, "Audio-visual evaluation of oratory skills," arXiv:2110.01367, 2021.