International Journal for Innovative Engineering and Management Research
PEER REVIEWED OPEN ACCESS INTERNATIONAL JOURNAL
www.ijiemr.org

AI-POWERED MULTIMODAL COMMUNICATION ANALYSIS FOR PERSONALIZED FEEDBACK

Mr. Sardar Prabhpreet Singh, Mr. Avinash Sharma, Mr. Bitla Reyansh

Undergraduate Students, BS-MS (Computer Science), Vishwa Vishwani Institute of Systems and Management
Hyderabad, Telangana, India | Email: sardarprabhpreetsingh@gmail.com | Mobile: 9515080634

ABSTRACT

Good communication is important in many areas of personal life, such as education and in terms of career development. Yet those who want to enhance their communication skills usually do not have the chance to get individual and practical feedback regarding their speaking style and non-verbal behaviour. The paper introduces an AI-driven multimodal communication coaching framework which allows people to find out where they need to improve by analysing both their speech and video. By examining verbal cues such as speaking rate, filler words, repetitions, and speech clarity, as well as non-verbal cues like posture, gestures, and head position, the proposed approach provides recommendations that are in line with the user's communication objectives. The various applications include interview preparation, public speaking, presentations, and ordinary workplace discussions. The study adopts a design-oriented method that is based on the current research literature concerning multimodal analysis and personalised feedback. Moreover, the paper looks at the technical limitations and ethical issues of the proposed framework, such as the effect of recording quality, differences in communication styles, privacy worries, and algorithmic bias. In general, the proposed solution is expected to assist conventional learning and coaching methods without taking away the human element. The effectiveness, accuracy, and suitability of the framework will have to be confirmed by carrying out a prototype and subjecting it to rigorous testing.

Keywords: Multimodal Artificial Intelligence, Communication Coaching, Speech Analysis, Computer Vision, Personalized Feedback, Responsible AI.

1. INTRODUCTION

Spoken communication is used in classroom discussions, interviews, presentations, teamwork, and routine conversations. Being understood depends not only on the words a person chooses but also on delivery: pace, organization, pauses, and repeated expressions can all affect how a message comes across. People preparing for a presentation or interview may not have access to a coach. Listening to recordings can help, but reviewing them carefully takes time and it may be hard to notice recurring habits. Speech recognition and generative AI offer one possible way to make this kind of practice easier to access. Automatic speech recognition converts an audio recording into text, and a program can calculate simple statistics such as word count or filler-word frequency. A generative model can then use the transcript and measurements to produce explanations or practice ideas. Whisper is a speech-recognition system trained with large-scale weak supervision (Radford et al., 2023), while Google provides the Gemini API for generating model responses (Google, n.d.). SpeakSense is designed to bring these steps together in a web application. A user submits a recording, the system transcribes it, calculates

selected measures, and requests feedback based on the available evidence. The planned scope is audio-only: the application does not currently assess eye contact, facial expression, posture, gestures, or emotional state. It is intended as a self-practice aid, not as a basis for high-stakes decisions. The contribution proposed here is an integrated workflow built from existing technologies. Its value should be judged by the reliability of its measurements, the relevance of its recommendations, and its performance on real recordings—not simply by the fact that it uses AI.

2. REVIEW OF LITERATURE

Radford et al. (2023) introduced Whisper, a speech-recognition system trained on large collections of audio and text. Their work examined multilingual recognition and reported performance across a range of speech-processing benchmarks. Whisper is the planned transcription component in SpeakSense. Because recognition mistakes can affect later counts and text-based analysis, users should be able to inspect the transcript rather than assume it is error-free. Ochoa and Zhao (2024) presented OpenOPAF, an open-source system for automated oral-presentation feedback. Its analysis includes features such as vocal volume, speaking speed, filled pauses, articulation, body language, gaze, and text shown in visual aids. This work demonstrates how automated systems can provide feedback on presentation performance. SpeakSense has a narrower initial scope: it works with audio and transcript-derived measures rather than video-based behaviour. Ochoa and Dominguez (2020) studied a multimodal feedback system in a semester-long learning activity involving 180 students. They reported that outcomes differed across presentation dimensions and found a small but statistically significant effect when later presentations were evaluated by a human expert. Their findings illustrate why a feedback system should be judged by outcomes and independent assessment, not only by whether it generates a report. Hummel et al. (2025) analysed 14 automated oral-presentation feedback systems across 83 functional features and other aspects. Their review identified areas for improvement, including feature integration, adaptive feedback, content structure, and the evaluation of learning outcomes. These observations support the need for careful design and practical testing. Overall, the literature indicates that speech recognition and automated feedback are feasible, while also showing that measurement quality and the effects of feedback need to be examined. SpeakSense does not introduce a new speech-recognition model. Instead, it proposes to combine transcription, lightweight quantitative analysis, and generated coaching suggestions in one web application.

3. RESEARCH GAP

Previous research and existing systems show that automated tools can examine selected features of speech and provide communication feedback. A remaining design challenge is to turn measurements into advice that people can understand and use during practice. A transcript records recognized words, but it does not by itself describe every aspect of delivery. In addition, a language model may produce advice that sounds plausible without being adequately supported by the recording. SpeakSense is designed to address this practical challenge by separating the calculation of metrics from the generation of feedback. Python calculates

predefined statistics, and Gemini receives the transcript and those measurements as context when producing recommendations. The intended interface presents the results in a straightforward way and distinguishes calculated values from AI-generated interpretation. This is the motivation for the project, not a claim that no comparable system exists. Any claim of novelty would need to be supported by a systematic comparison with current tools and published research.

4. OBJECTIVES OF THE STUDY

- To develop a web application that accepts voice recordings for communication practice.
- To transcribe submitted speech with Whisper and display the resulting transcript.
- To calculate selected speech measures, including word count, estimated speaking rate, filler-word frequency, and repeated expressions.
- To generate recommendations through the Gemini API using the transcript and calculated measures as context.
- To present results in a clear dashboard and preserve basic transcript and metric functionality when the AI feedback service is unavailable.
- To define a testing approach for technical reliability and the quality of generated feedback.

5. RESEARCH METHODOLOGY

Research design: This paper follows a descriptive, literature-informed system-design approach. It reviews selected research and technical documentation and explains a proposed implementation. It does not report a completed user study or experimental results.

Sources and intended data: The literature review draws on research about speech recognition and automated communication feedback, along with technical documentation for the planned software components. After implementation, the application is intended to process audio recordings supplied by users.

System workflow: A user uploads an audio file or records speech through a supported browser interface. The FastAPI backend validates the file and prepares it for processing. Whisper generates a transcript. Python calculates selected statistics from the transcript and, where supported, from audio timing or audio analysis. The backend sends the transcript and calculated measures to Gemini, and the frontend displays the transcript, measurements, and recommendations.

Technology stack: The proposed frontend uses React and Vite, with CSS for styling. Python and FastAPI provide backend endpoints; Whisper handles transcription; Python and suitable audioprocessing libraries support metric calculation; and the Gemini API generates contextual feedback. REST endpoints connect the frontend and backend. API credentials should remain on the server and must not be exposed in browser code.

Speech measures: Word count should use a consistent tokenization rule. Speaking rate may be estimated as the number of recognized words divided by the selected duration in seconds, multiplied by 60. The application should state whether this duration covers the full recording or only speech-active time. Filler-word detection depends on a defined list of expressions, while repetition analysis looks for configured repeated words or short phrases. Pause information should be shown only when reliable timestamps or audio-based methods are available. These measures are estimates and can be affected by recognition errors.

AI-generated feedback: Gemini receives the transcript and calculated measures as separate inputs. The prompt should ask for evidence-linked strengths, possible areas to practise, specific suggestions, and a short exercise. It should not invent numerical results or infer confidence, emotion, honesty, or personality from the transcript. Outputs should be checked before display. If

the AI service is unavailable, the application should still show the transcript and any metrics that were successfully calculated.

Evaluation plan: After implementation, functional testing should cover valid and invalid files, transcription, metric calculations, service failures, and result display. A set of manually checked transcripts can be used to estimate word error rate. Filler-word and repetition counts can be compared with manual annotations. Human reviewers can assess feedback for correctness, relevance, clarity, and actionability. The test set, procedure, and results should be documented. No performance figures should be reported before the tests have been completed.

6. CONCEPTUAL BACKGROUND

Automatic speech recognition (ASR): ASR converts an audio signal into text. In SpeakSense, Whisper provides the transcript used for later analysis. Recognition errors can affect every downstream metric based on that text.

Quantitative speech analysis: Quantitative speech analysis calculates defined features from audio or transcripts. Planned examples include word count, estimated speaking rate, filler-word counts, repetitions, and pauses when dependable measurements are available. These features describe specific patterns; they do not independently establish overall communication quality.

Generative AI feedback: Generative AI produces natural-language responses from supplied context. SpeakSense uses Gemini to turn the transcript and calculated measures into explanations and practice suggestions. Because generated statements can be unsupported or inaccurate, feedback needs checking.

Personalized communication practice: Personalized practice links recommendations to a particular recording and, where possible, the user's goal. For example, if a recording contains frequent detected filler words, the system might suggest practising deliberate pauses. Advice should refer to the observed pattern rather than make a broad judgment about the speaker.

7. ANALYSIS AND DISCUSSION

7.1 Pipeline integration: SpeakSense brings transcription, programmatic measurements, and AIgenerated feedback into one workflow. The separation of responsibilities makes it easier to distinguish numerical outputs from the model's explanation and can help developers identify where an error occurred.

7.2 Reliability of speech measures: The dashboard is only as useful as the definitions and calculations behind it. Speaking rate may be misleading when the duration includes long periods of silence. Filler-word counts can be distorted by recognition errors or by words that serve a legitimate purpose in context. Repeated phrases may also be intentional. The interface should explain how each measure is calculated and avoid labelling every detected pattern as a problem.

7.3 Quality of recommendations: A language model can phrase feedback clearly, but its suggestions may still be generic or unsupported. Providing the transcript and measured statistics as explicit context may help anchor the response, but it does not guarantee correctness. A sample of outputs should be reviewed by people before conclusions are drawn about feedback quality.

7.4 Privacy and responsible use: Recordings and transcripts may contain personal information. The application should explain how data is processed, restrict access, protect API credentials, and avoid retaining audio longer than necessary. Users should be informed when an external service processes their data. SpeakSense is intended for practice and should not be presented as a validated tool for hiring, grading, clinical diagnosis, or psychological assessment.

7.5 Scope and practicality: Limiting the first version to audio makes the project more manageable than a system that analyses both audio and video. The planned system can report selected transcriptbased patterns, but it cannot assess visual communication. Pronunciation accuracy, emotion, confidence, and voice quality should not be described as measured capabilities unless specific methods are implemented and validated.

7.6 Need for evaluation: Prior studies show that the impact of automated feedback can differacross presentation skills, and reviews have identified limited evaluation of learning outcomes in some systems (Ochoa & Dominguez, 2020; Hummel et al., 2025). SpeakSense should therefore be tested for successful processing, measurement accuracy, and whether its recommendations are evidencebased and useful.

8. MAJOR FINDINGS

The following points are drawn from the reviewed literature and the proposed design. They are not experimental findings from a completed SpeakSense deployment.

- Research supports speech recognition as a possible basis for analysing spoken content (Radford etal., 2023).
- Automated feedback systems can examine selected communication features, although their coverage and functions differ (Ochoa & Zhao, 2024; Hummel et al., 2025).
- Automated coaching should be evaluated using learning-related outcomes and independent assessment, rather than the existence of a generated report alone (Ochoa & Dominguez, 2020).
- Keeping calculated metrics separate from AI-generated explanations makes it easier to inspect and verify the system's outputs.
- Transcription mistakes, ambiguous filler words, recording duration, and unreliable timestamps may reduce the accuracy of speech statistics.
- Real-recording tests and human review are needed before claims can be made about SpeakSense's accuracy, usefulness, or ability to improve communication.

9. SUGGESTIONS AND RECOMMENDATIONS

Test the application on a varied set of real recordings and document the conditions and outcomes.

- Check speech measurements against manually reviewed transcripts or annotations before reporting accuracy.
- Provide Gemini with the transcript and calculated measures, and instruct it to avoid inventing numbers or unsupported observations.
- Explain each metric in the interface, including how speaking duration and filler-word frequency are calculated.
- Keep transcription and available basic measurements usable when the Gemini API is unavailable.

- Protect recordings and transcripts, keep API keys on the backend, and clearly describe data-retention practices.
- Do not claim to measure confidence, emotion, pronunciation accuracy, or overall communication ability unless the relevant methods have been implemented and validated.
- For future work, consider progress tracking across multiple sessions and user studies comparing feedback-supported practice with an appropriate comparison condition.

10. CONCLUSION

This paper has described SpeakSense, a proposed web application for voice-based communication practice. Its planned workflow combines Whisper transcription, Python-based speech measures, Gemini-generated feedback, and a React frontend connected to a FastAPI backend. Expected outputs include a transcript, word count, estimated speaking rate, filler-word and repetition analysis, pause information where dependable, and practical recommendations. The literature reviewed here indicates that automated speech processing and communication feedback are feasible, while also highlighting the need to test measurement reliability and feedback quality. SpeakSense's proposed contribution is the integration of

existing tools into a workflow that separates calculated statistics from AI-generated explanations. The system's accuracy and practical value remain unverified until it has been tested with real recordings. Future evaluation should examine transcription and metric reliability, have reviewers assess the recommendations, and investigate whether repeated use supports improvement in spoken communication.

REFERENCES

Google. (n.d.). Generate content with the Gemini API. Google AI for Developers. https://ai.google.dev/api/generate-content

Hummel, S., Schneider, J., Mouhammad, N., Klemke, R., & Di Mitri, D. (2025). Enhancing presentation skills: Key technical features of automated feedback systems—A systematic feature analysis. International Journal of Technology Enhanced Learning, 17(6). https://doi.org/10.1504/IJTEL.2025.148593

Ochoa, X., & Dominguez, F. (2020). Controlled evaluation of a multimodal system to improve oral presentation skills in a real learning setting. British Journal of Educational Technology, 51(5), 1615–1630. https://doi.org/10.1111/bjet.12987

Ochoa, X., & Zhao, H. (2024). OpenOPAF: An open source multimodal system for automated feedback for oral presentations. Journal of Learning Analytics, 11(3), 224–248. https://doi.org/10.18608/jla.2024.8411

OpenAI. (n.d.). Whisper [Computer software]. GitHub. https://github.com/openai/whisper
Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., & Sutskever, I. (2023). Robust speech recognition via large-scale weak supervision. In Proceedings of the 40th International Conference on Machine Learning (Vol. 202, pp. 28492–28518). Proceedings of Machine Learning Research. https://proceedings.mlr.press/v202/radford23a.html