2026 ASEE ANNUAL CONFERENCE June 21–24, 2026 | Charlotte Convention Center, NC
Host Sponsor UNC Charlotte
ASEE AMERICAN SOCIETY FOR ENGINEERING EDUCATION
Paper ID #50273

Work in Progress: Beyond Words – A Multimodal AI Coaching Tool to Enhance Communication Skill Development in Engineering Education

Jeslyn Wang, University of Toronto

Jeslyn Wang is a Computer Engineering student at the University of Toronto specializing in artificial intelligence. Her research focuses on multimodal learning systems and AI-driven educational tools. She previously worked in applied machine learning and is passionate about developing technology that enhances communication and learning.

Eren Cimentepe, University of Toronto
Guang Yang, University of Toronto

A fourth-year Computer Engineering student at the University of Toronto, pursuing minors in Engineering Business and Artificial Intelligence.

Yu Ze Zheng
Dr. Hamid S Timorabadi P.Eng., University of Toronto

Hamid Timorabadi received his B.Sc, M.A.Sc, and Ph.D. degrees in Electrical Engineering from the University of Toronto. He has worked as a project, design, and test engineer as well as a consultant to industry. His research interests include the applications in Engineering Education.

Beyond Words – A Multimodal AI Coaching Tool to Enhance Communication Skill Development in Engineering Education

Abstract

Effective communication relies on both verbal and non-verbal cues, yet existing coaching tools primarily emphasize audio feedback and fail to capture critical visual and contextual dimensions such as facial expressiveness, emotions, and scenario-specific delivery. As a result, current systems lack an integrated, context-aware approach to holistic communication training.

This study evaluates an educational intervention aimed at addressing these gaps by developing an AI-powered communication coaching assistant that integrates video, audio, and spoken content analysis to deliver comprehensive, context-aware feedback. By combining speech recognition, computer vision, and natural language processing, the system provides personalized and actionable guidance on vocal delivery, facial expression, body language, and speech coherence. Its contextual awareness enables adaptive feedback for specific scenarios, such as technical interviews or classroom presentations, making it directly relevant to engineering education.

To validate its effectiveness, the system was evaluated using a pilot user study design informed by prior peer-reviewed research in communication training. Post-study survey results indicated that approximately 93% of participants reported improvement in their presentation and communication skills after using the platform and receiving feedback, demonstrating functional reliability and meaningful educational impact.

The intended contribution is a holistic, accessible communication training tool for engineering students, job seekers, and professionals who may lack access to in-person mentors or to costly coaching. The paper presents the system's technical design, multimodal AI integration, evaluation methodology, and early user study findings, and discusses its potential role in supporting scalable, effective communication improvement across educational and professional contexts.

1.0 Introduction

Good communication skills are essential across many educational, professional, and social contexts. Whether in academic settings, job interviews, collaborative team discussions, or public-facing roles, individuals must articulate their ideas confidently and persuasively. Effective communication is consistently linked to academic success, employability, and good leadership [1] [2] [3]. Yet, despite its importance, structured training often remains insufficient and unequal across individuals due to challenges in incorporating it into every degree [4].

While effective communication is inherently multimodal, combining verbal content, facial expressions, body language, and vocal delivery, formal education remains largely content-focused [5], [6]. This creates a mismatch between academic training and real-world professional expectations [7]. This gap has been widened by the rise of remote learning and virtual interactions, which have reduced opportunities for students to develop these skills through in-person feedback and led to a measurable decrease in communication confidence [8][9][10]. Furthermore, although traditional coaching and mock interviews can bridge this gap [11], they are often too costly or time-intensive for many learners, particularly those from underrepresented or lower-income backgrounds [12][13]. Consequently, many students are forced to navigate high-stakes presentations and interviews with minimal personalized guidance.

In response to these systemic gaps, this paper presents a multimodal AI-powered educational intervention specifically designed to scale communication coaching within engineering curricula. While general-purpose tools often prioritize emotional expressiveness or vocal variety, this system is designed for technical communication where clarity, structured logic, and professional neutrality are the primary metrics of success. The proposed pipeline extracts synchronized features from video, audio, and transcripts. It uses unsupervised clustering to identify recurring behavioral patterns and leverages a Large Language Model (LLM) to aggregate these signals into context-specific feedback tailored to engineering scenarios such as capstone reviews or technical interviews. By centering the design on the unique rhetorical requirements of the engineering profession, the system provides a scalable alternative to resource-intensive human mentorship.

Unlike existing tools that primarily target one modality in depth without considering the bigger picture, the new approach will aim to evaluate communication that more closely approximates the real world. The remainder of this paper will describe the system architecture, design choices, the multimodal feature extraction and clustering methodology, and the integration of semantic analysis for content-level feedback. The evaluation approach, including preliminary testing and results, will then be presented. Finally, the educational implications, limitations, and future directions for how multimodal AI systems can enhance communication skill development and be used in grading applications will be discussed. The primary contribution of this work is an application-specific AI framework that addresses the scalability constraints of communication training in engineering curricula.

2.0 Related Work

This section emphasizes the project's position at the intersection of communication skills training and education, outlining existing AI-based coaching tools and emerging multimodal approaches to feedback. While the suggested approaches have broad applications, a substantial portion of professional communication skill development occurs during formal education and training [14] [15]. Consequently, the following review evaluates existing coaching tools and emerging multimodal methodologies designed to improve these critical skills.

2.1 Communication Skills Training in Education

Many existing educational tools that help individuals develop communication skills rely on instructor feedback (e.g., professors and supervisors), peer review, and workshops [14]. These methods have been shown to improve many students' clarity, organization, and delivery through iterative practice and critique. However, these methods rely on finite instructor time and are difficult to scale as student populations grow, often resulting in high-level overviews rather than the specific improvements necessary for skill development. This is particularly evident in engineering programs where individualized feedback is limited, and peer review can be inconsistent in quality [16]. Consequently, many students receive sparse, delayed, and non-actionable feedback that fails to support meaningful growth. This gap has motivated growing interest in technology-supported feedback systems that can provide immediate, individualized, and repeatable coaching outside the classroom.

2.2 Existing AI-Based Coaching Tools

Commercial and research systems such as Orai [17], Yoodli [18], and Google's Interview Warmup [19] analyze features such as speech rate, filler word frequency, pauses, and prosodic metrics, including pitch and volume, to provide surface-level delivery feedback.

In academic research, interview coaching systems have focused on analyzing the speaker's audio and vocal delivery metrics, with limited integration of visual behaviour or facial cues into their feedback mechanisms [20]. Although useful for improving aspects of speech, these audio-centric approaches do not provide a holistic view of communication, as nonverbal cues strongly influence how a presentation or speech is received by an audience [21]. Furthermore, feedback from many existing tools remains generic and context-independent, offering similar suggestions regardless of whether the user is preparing for a technical interview, a behavioural interview, or an academic presentation [20]. Due to these limitations, a holistic system that integrates multiple modalities and adapts feedback to the communicative context is necessary.

2.3 Multimodality and Machine Learning in Education

Early multimodal communication coaching systems, such as MACH, an automated conversation coach [22], and Presentation Trainer [23], demonstrate the value of combining audio and visual cues for interview and public speaking training. MACH integrates facial expression analysis with prosodic speech features to provide feedback in job interview scenarios, while Presentation Trainer tracks vocal delivery and body movement to support real-time coaching for public speaking. While these systems highlight the effectiveness of multimodal delivery feedback, they primarily focus on nonverbal behaviour and acoustic features for real-time coaching.

Unsupervised learning techniques, such as clustering of behavioural features, have been used to identify engagement patterns and recurring speaking styles such as charisma without the need for labelled data [24]. This is particularly relevant in interview and presentation coaching contexts, where behavioural labels are subjective and large, annotated datasets are difficult to obtain. Clustering enables the discovery of delivery patterns using extracted features alone.

Simultaneously, there have been several advances in LLMs, enabling research on API-based analysis of speech components such as coherence, relevance, and sentence structure [25].

These experiments highlight LLMs' ability to evaluate speeches and the ways they are communicated, including their organization and meaning.

Despite the progress in this area, there are relatively few systems that incorporate all three of audio, visual, and semantic information with machine learning assistance, especially in the context of engineering education. Additionally, the feedback generated is not sufficiently context-specific or flexible, focusing either on presentation or interview practice and not accommodating the different contexts in which effective communication is required. This gap again motivates the multimodal approach presented in this paper, which provides more holistic, situation-specific feedback to users.

3. Methodology

We developed a multimodal pipeline that generates automated oral-communication feedback from presentation videos. Users upload an MP4 and provide a brief context label (e.g., capstone presentation, mock interview). The system extracts synchronized audio, visual, and transcript streams and produces timestamped, context-aware feedback (Figure 1).

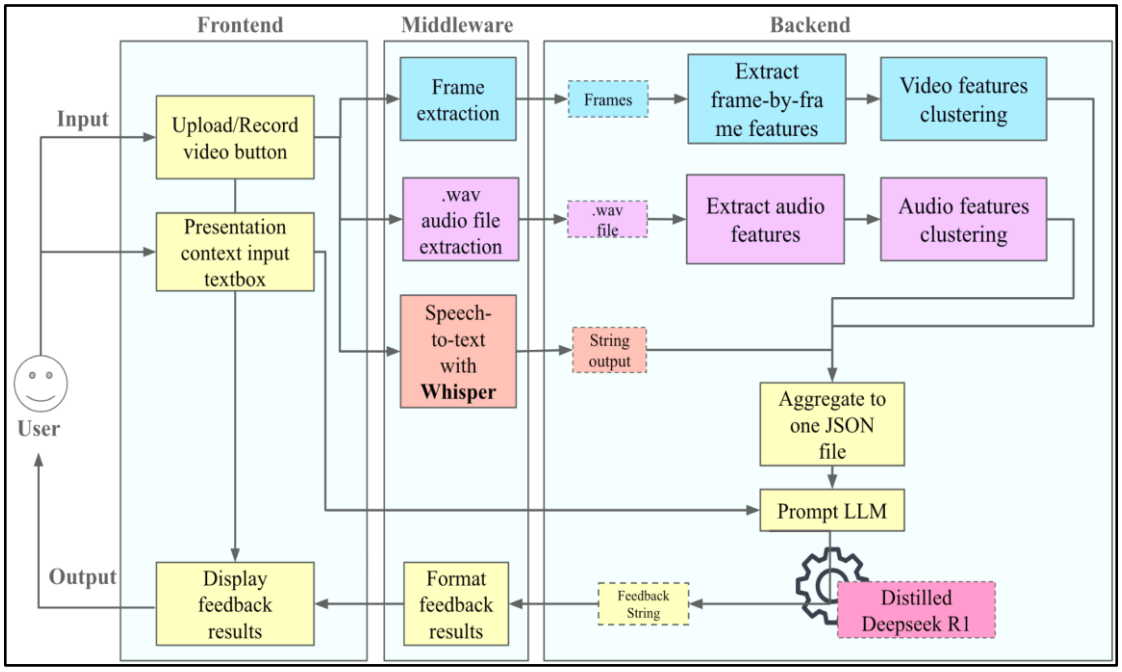

Figure 1. Multimodal LLM-powered system leveraging video, audio, and speech-to-text data to generate an enhanced evaluation of a user's oral presentation.

3.1 Data Input and Segmentation

Each uploaded video is decomposed into three synchronized streams: audio, visual frames, and spoken content. All streams are segmented using shared timestamps to ensure temporal

alignment across modalities. This unified segmentation enables integrated multimodal analysis at the segment level.

3.2 Audio Analysis

Each audio segment is analyzed based on three complementary dimensions: vocal delivery (clarity and fluency), perceived confidence, and emotion. For each dimension, a dedicated feature subset is extracted using OpenSMILE. Unsupervised K-means clustering is then applied within each feature space to identify recurring vocal behaviour patterns. As a result, each audio segment is assigned a cluster membership within each model, corresponding to its delivery, confidence, and emotional profiles.

3.3 Visual Analysis

The visual analysis pipeline objectively quantifies nonverbal behaviour by processing video frames using the Python Facial Expression Analysis Toolbox (Py-Feat). The system extracts granular Facial Action Units (AUs) along with probability scores for seven standard emotional states. To convert this raw data into interpretable feedback, features are aggregated into 3-second windows and subjected to K-means clustering, which categorizes the speaker's performance into discrete behavioural states (e.g., "dynamic engagement" vs. "static neutrality").

3.4 Language and Content Analysis

While the audio and visual pipelines capture how information is delivered, the language and factual analysis component focuses on what is being communicated and how effectively it is structured. This module evaluates the coherence, clarity, relevance, and correctness of the user’s response. To achieve this, the user input transcript is extracted from the generated audio segments. This enables the transcripts to align with the corresponding audio and visual segments and to be incorporated into the LLM's multimodal input.

3.5 Multimodal Integration and Feedback Generation

Audio cluster labels, visual behavioural states, and transcript segments are merged into a unified JSON representation using timestamp-based concatenation. This multimodal data, combined with the user-specified presentation context, is provided to an external LLM (DeepSeek R1 Qwen3 8B) via API. A key contribution of this system is the Engineering Context Module. By allowing users to specify a context (e.g., 'Technical Interview' or

'Capstone Design Review'), the backend LLM (DeepSeek R1) adjusts its weighting heuristics. For instance, in a technical context, the system is tuned to be less penalizing toward 'professional neutrality' (often misclassified as monotone by general tools) and focuses instead on the precise alignment of verbal technical explanations with non-verbal confidence cues. The LLM synthesizes multimodal signals to generate timestamped, context-specific feedback that reflects both delivery and content effectiveness.

3.6 Model Selection Rationale

To ensure the system's reliability, underlying models were selected based on established industry and academic standards. We used OpenAI Whisper for transcription because of its demonstrated effectiveness in handling specialized terminology and vocabulary for specific topics, accents, and spontaneous speech, including filler words and hesitations, which are highly prevalent in engineering communication contexts [26][27].

For acoustic feature extraction, OpenSMILE was selected because it remains the industry standard for paralinguistic analysis, particularly through its eGeMAPS feature set, which has been validated in affective computing and behavioural signal processing research [28]. This ensures that our confidence and vocal quality metrics are grounded in established signal-processing literature rather than in experimental random effects [29].

Lastly, for visual feature extraction, Py-Feat was chosen for its validated implementation of Facial Action Coding System (FACS)-based action unit detection, enabling interpretable behavioural analysis aligned with established psychological frameworks [30], which are effective in prior work.

4.0 Educational Validation and Learning Design

The primary objective of this validation is to assess the system's effectiveness as a formative educational tool for developing engineering communication skills, rather than to evaluate technical performance alone. Communication proficiency is learned through iterative practice, feedback, reflection, and revision, yet engineering curricula often lack scalable mechanisms to support this process. The system is therefore evaluated based on its ability to facilitate a structured learning loop in which students produce an initial response, receive multimodal, timestamped feedback, and apply that feedback to a revised performance. This approach directly targets core communication competencies emphasized in engineering education, including clarity, organization, delivery, and professional presence.

The validation framework is grounded in formative assessment and self-regulated learning, emphasizing learner awareness and agency over summative scoring. Educational effectiveness is measured by student-reported usefulness, gains in confidence, and perceived improvement, reflecting whether learners can meaningfully interpret feedback and translate it into observable performance changes. By enabling repeated, low-stakes rehearsal with immediate, actionable feedback, the system addresses common instructional constraints such as limited instructor time and uneven access to individualized coaching. As a result, the platform functions as a scalable educational resource that complements traditional instruction while expanding opportunities for reflective communication practice in engineering education.

4.1 Study Participants & Recording Protocol

To evaluate the system's functional performance and usability, a pilot study was conducted using convenience sampling with 15 engineering students. This sample size is appropriate for a pilot usability study , where the objective is to verify functional behaviour and collect preliminary user feedback rather than achieve statistical generalization. Each participant provided consent to record two short videos for research and evaluation purposes. To isolate the system's performance and ensure comparability across test runs, all recordings followed standardized environmental controls: an indoor setting with minimal background noise, an eye-level camera capturing the face and upper body, the exclusion of any visual aids, and a strict 1-2 minute duration.

4.2 Experimental Design

Within-Subjects Structure

In within-subject experiments, all participants experience every condition. So, in this experiment, each participant completed two recordings with the following prompt: "*Explain how search engines choose which results to show"*. The design details are outlined in Table 1.

Table 1. Within-Subjects Experimental Conditions and Evaluation Objectives

Condition	Description	Purpose
A — No Preparation	Participants immediately recorded a 1–2 minute video in response to the prompt.	Establishes baseline clarity, structure, communication style, and topic knowledge.
B — With Preparation	Participants reviewed the AI-generated feedback from their baseline recording alongside a one-page written explanation of the topic. After reviewing both, they recorded a second 1–2 minute video answering the same question.	Measures improvement in organization, clarity, accuracy, and confidence after receiving background knowledge.

Comparing Condition A vs. B enables direct measurement of user improvement. If users show measurable improvements in pacing, clarity, filler reduction, or confidence, this indicates that the tool provides meaningful, actionable feedback rather than generic comments.

4.3 Blind Expert Assessment and Rubric Design

To address the limitations inherent in self-reported data and measure actual skill acquisition, the study employed a rigorous blind review process. A panel of four independent annotators, consisting of engineering student peers, evaluated the anonymized video pairs. To mitigate confirmation bias, annotators were blinded to the experimental condition and viewed recordings in a randomized order, unaware of which video represented the baseline or post-intervention attempt. Each submission was scored using a standardized 20-point rubric derived from the faculty's oral presentation guidelines. This instrument assessed four distinct dimensions of engineering communication on a 5-point Likert scale: vocal delivery, nonverbal cues, technical clarity, and overall professionalism. Finally, the consistency of these subjective scores was statistically validated using the Intraclass Correlation Coefficient (ICC) to confirm inter-rater reliability.

4.4 User Feedback Measures

After completing both recordings, participants filled out a short survey assessing:

- Ease of use of the tool (1–10)
- Usefulness of feedback (1–10)
- Accuracy of comments (1–10)
- Whether feedback helped them improve on the second recording
- Likelihood of using the tool again
- Open-ended suggestions or issues encountered

Survey responses provide quantitative measures of usability and qualitative insights into user experience.

5.0 User Study and Early Findings

Following the experimental protocol and evaluation metrics outlined in Section 4, the pilot study was conducted with 15 participants (n=15). The resulting data was analyzed across objective expert assessments, system accuracy benchmarks, and subjective user perception surveys.

5.1 Evaluation Metrics

To address the limitations of self-reported data, performance was evaluated through a blind expert assessment and a user-experience survey. To measure actual skill acquisition, we employed a blind review process. A panel of four independent annotators (engineering teaching assistants and peers) scored the anonymized video pairs. To prevent bias, annotators were blinded to the condition (i.e., they did not know which video was the "before" or "after" attempt). Each video was scored using a 20-point rubric, assessing vocal delivery, nonverbal cues, technical clarity, and professionalism on a five-point scale. For the user experience survey, participants completed a post-study questionnaire rating the system on usefulness, accuracy, and ease of use (1–10 scale) and provided qualitative feedback on specific features.

5.2 Results and Interpretation

5.2.1 Objective Performance Gains

Analysis of the blind expert scores revealed a measurable improvement in student performance following the AI feedback intervention. The mean score rose from 12.0 in the Baseline condition to 15.8 in the Post-Intervention condition, representing a 31.7% increase in overall presentation quality as shown in Figure 2. The most significant improvement was observed in overall professionalism, with scores increasing by 40.0%. This holistic metric evaluates communication readiness for professional settings like engineering interviews. This

substantial increase demonstrates that the system's multimodal feedback successfully helped participants self-correct distracting habits, elevating their delivery to a standard the human scorers deemed more prepared for a professional technical environment.

To establish the reliability of the blind expert assessments, a two-way mixed-effects model was used to calculate the Intraclass Correlation Coefficient (ICC2) for absolute agreement. The resulting ICC for the overall presentation score was 0.81, indicating good inter-rater reliability among the three annotators [31].

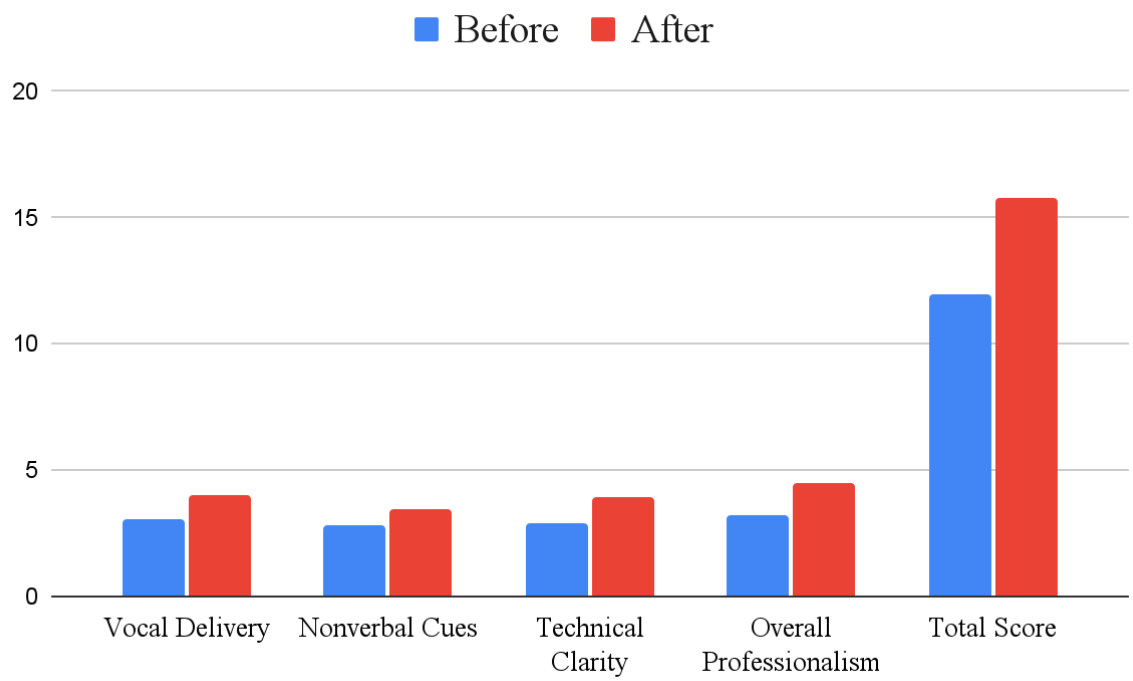

Figure 2. Average scores of participants (n=15) from four independent annotators before and after receiving feedback from the multimodal AI tool.

5.2.2 User Perception and Confidence

Quantitative results indicate high user engagement and strong perceived value, particularly in the system's ability to facilitate self-reflection. A critical requirement for scalable educational tools is a low technical barrier to entry. As shown in Figure 3, participants rated the ease of the video upload and feedback retrieval workflow exceptionally high, yielding a mean score of 9.47 out of 10. This confirms that the system's frontend architecture and data ingestion pipeline do not impede the learning process. Participants were also asked to evaluate the AI-generated insights. Figure 4 illustrates the distribution of scores for the usefulness of the feedback, with a mean score of 7.67 out of 10. Similarly, when asked to rate the accuracy

with which the feedback reflected their actual delivery and spoken content, participants reported a mean accuracy score of 7.67 (Figure 5). These scores suggest that, although the multimodal analysis is generally well-received and closely aligns with students' self-perceptions, there remains room to further calibrate the LLM's contextual sensitivity.

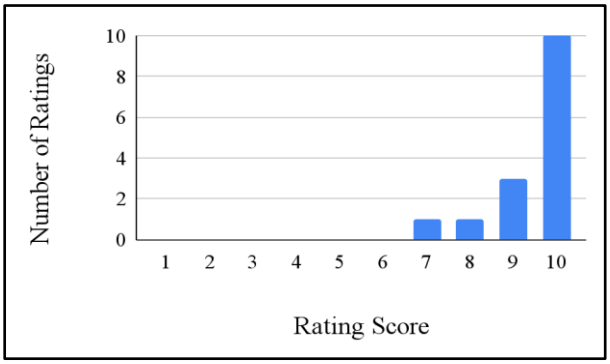

Figure 3. Distribution of participant ratings for system usability (n=15). Participants rated the ease of the upload and feedback workflow on a scale of 1 (not easy) to 10 (very easy), yielding a mean score of 9.47.

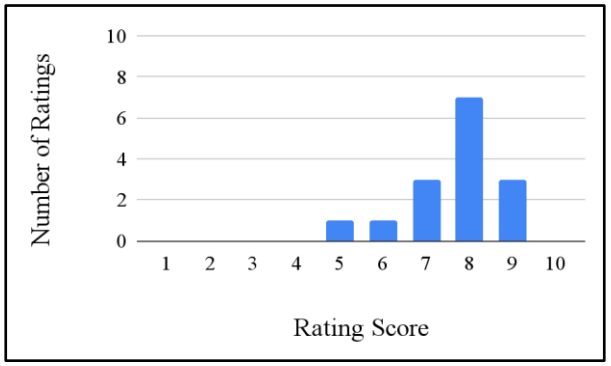

Figure 4. Distribution of participant ratings for usefulness of generated feedback (n=15). Participants rated the ease of the upload and feedback workflow on a scale of 1 (not useful) to 10 (very useful), yielding a mean score of 7.67.

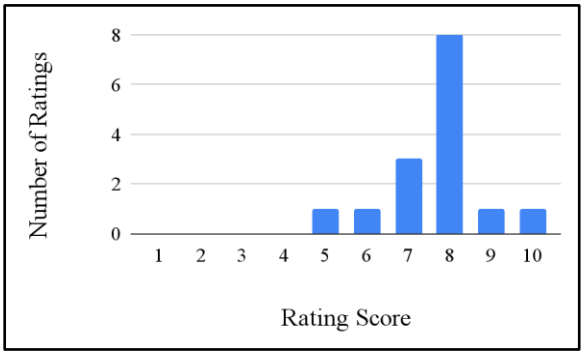

Figure 5. Distribution of participant ratings for generated feedback accuracy (n=15). Participants rated the ease of the upload and feedback workflow on a scale of 1 (not accurate) to 10 (very accurate), yielding a mean score of 7.67.

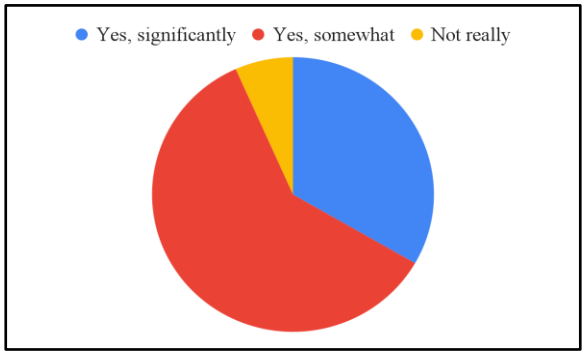

Figure 6. Participant self-reported improvement (n=15). 60% selected “Yes, somewhat”, 33% selected “Yes, significantly”, and 7% selected “Not really”.

These quantitative results show that users found the process overall frictionless, that the majority found the insights valuable, and that the multimodal system’s feedback aligned with their own self-perceptions. Most notably, 93% of participants (14/15) reported that the AI-generated feedback helped them improve their second recording, with 33% stating it helped "significantly," as shown in Figure 6.

Analysis of open-ended feedback revealed participants’ insights regarding the system's utility and areas for refinement. Participants called for greater contextual nuance, noting that standard critiques, such as flagging "monotone delivery", were occasionally misaligned with the expectations of technical professional settings. In addition, participants validated the

importance of granular evidence by requesting enhanced, actionable visualizations, specifically clickable timestamps that directly link feedback to video segments. Finally, participants noted that numerical scores did not always align with perceived qualitative improvement, suggesting that the scoring heuristic requires calibration.

5.3 Interpretation

The high reported improvement (93%) despite the moderate accuracy ratings (7.6/10) suggests that the multimodal feedback loop itself (e,g., recording, receiving objective data, and iterating) is instructionally valuable, even when the AI's specific semantic judgments are imperfect.

The qualitative feedback regarding generic tone criticism specifically informs future work. It indicates that students want advice tailored to their specific rhetorical situation (e.g., technical interview vs. storytelling). The results confirm that while clustering visual and audio features is technically successful, the interpretation of those clusters must be dynamically adjusted based on the user's stated intent to avoid "false positive" critiques of professional neutrality.

6.0 Discussion

Our work examines whether a multimodal, context-aware coaching system can provide scalable, actionable feedback to engineering students preparing for presentations and interview-style communication [14] [15]. Because our system combines visual, vocal, and language-level cues from the user, it more accurately reflects how communication is assessed in authentic engineering communication settings than audio-only or single-metric tools [5] [6]. Although our user results are preliminary, they suggest the system supports repeated low-stakes rehearsal by giving immediate feedback without instructor review.

6.1 Use Cases & Relationship to Traditional Feedback

The primary educational value is enabling repeatable, actionable, formative feedback. Since instructor or TA review time is limited, students often receive delivery-focused feedback only on final submissions [4]. In contrast, our system generates timestamped feedback after each rehearsal, pinpointing specific segments that need improvement, such as rapid pacing during technical explanations, extended pauses, reduced engagement cues, or unclear transitions. Context-aware weighting helps ensure feedback focuses on the behaviours most relevant to the scenario rather than applying a single, generic standard [4].

This application is designed to supplement, rather than replace, traditional instructor and collaborative feedback. Human feedback remains essential for discipline-specific expectations, nuanced coaching, and mentorship, especially when communication issues overlap with knowledge gaps or anxiety. Automated multimodal feedback can complement human review by delivering instant, consistent observations that enable more frequent practice, which is important for skill development [15]. By linking specific feedback to timestamps, such as detecting a drop in eye contact when a filler word is used, the tool offers a level of granularity that human observers often miss in real time. This data-informed technique enables students to self-correct technical aspects such as pacing, gaze, and posture independently, freeing instructors' time for higher-level concerns, such as technical content and organization, during course assessments.

6.2 Limitations and Threats to Validity

While our preliminary evaluation is promising, several limitations constrain generalizability and should be considered when interpreting results. Notably, our outcomes primarily reflect short-term, self-reported improvement; future work will include rubric-based performance measures and longitudinal tracking to better establish learning gains. First, our study is in its early stages, with a limited sample and a narrow set of presentation contexts, so the findings may not generalize to broader student populations. Second, model outputs can be sensitive to input conditions and population diversity. Clustering patterns may shift across different speaking styles, accents, and delivery habits, and multimodal features can degrade under varying lighting, camera angles, microphone quality, and background noise. Third, the tool’s evaluation may vary across runs and may not always reflect technical correctness or rubric intent without detailed fine-tuning.

7.0 Future Work

Our future work will focus on validating the learning impact in authentic engineering education settings and improving our tool’s overall performance/reliability. We plan to conduct more longitudinal studies that measure sustained skill development across repeated practice sessions, including pre- and post-comparisons using rubric-based scoring and user self-efficacy reflections. Studying this area can test whether multimodal feedback leads to immediate improvements and to transferable communication skills in subsequent presentations. A primary goal of ours is to eventually deploy and test the tool in a broader classroom setting. On the system side, we will strengthen fairness and consistency across

diverse speakers and conditions. We will evaluate performance across varied accents, cultural norms, and presentation styles, and incorporate rubric-anchored prompting, repetition tests, and consistency checks to reduce run-to-run variability and improve interpretability of feedback and scores.

8.0 Conclusion

Engineering students increasingly need strong communication skills for capstone design reviews, team collaboration, and professional interviews, yet scalable, individualized coaching is difficult to provide in large classes and remote or hybrid environments. This paper focuses on that gap by presenting a multimodal, context-aware AI coaching and feedback system that integrates visual, audio, and language-level signals to generate timestamped, concrete suggestions. In early benchmarking and user evaluations, the system demonstrated its ability to align automated feedback with rubric-relevant communication dimensions while keeping feedback concrete and actionable. Overall, this project offers an education-oriented approach to AI-assisted communication training and motivates larger, longer-term studies to evaluate learning gains, reliability across speaking styles, and practical use within engineering curricula.

References

[1] N. R. Salikhova, O. V. Grigoryeva, G. G. Semenova-Poliakh, A. B. Salikhova, O. V. Smirnikova, and S. M. Sopun, “Communication tools and social media usage: Assessing self-perceived communication competence,” Online Journal of Communication and Media Technologies, vol. 13, no. 4, e202343, 2023, doi: 10.30935/ojcmt/13453.

[2] H. Tushar and N. Sooraksa, “Global employability skills in the 21st century workplace: A semi-systematic literature review,” Heliyon, vol. 9, no. 11, art. e21023, 2023, doi: 10.1016/j.heliyon.2023.e21023.

[3] M. Erbay, M. S. Javed, J. C. Nelson, S. Benzerroug, E. A. Karkkulainen, and C. E. Enriquez, “The relationship between leadership and communication, and the significance of efficient communication in online learning,” Educational Administration: Theory and Practice, vol. 30, no. 6, pp. 2065–2076, 2024, doi: 10.53555/kuey.v30i6.5650.

[4] S. Johnson, S. Veitch, and S. Dewiyanti, “A framework to embed communication skills across the curriculum: A design-based research approach,” Journal of University Teaching & Learning Practice, vol. 12, no. 4, Art. 6, 2015.

[5] V. M. Silva, J. Holler, A. Özyürek, and S. G. Roberts, “Multimodality and the origin of a novel communication system in face-to-face interaction,” Royal Society Open Science, vol. 7, no. 1, p. 182056, 2020, doi: 10.1098/rsos.182056.

[6] E. Midgette, O. G. Stewart, and I. August, “Understanding multimodal assessment practices in higher education to improve equity,” Education Sciences, vol. 15, no. 11, art. 1523, 2025, doi: 10.3390/educsci15111523.

[7] S. P. Dewi, C. Gilligan, and A. Wilson, “Do the teaching, practice and assessment of clinical communication skills align across learning environments? An observational study,” BMC Medical Education, vol. 24, art. 609, 2024, doi: 10.1186/s12909-024-05596-8.

[8] H. Baber, “Social interaction and effectiveness of online learning: A moderating role of maintaining social distance during the COVID-19 pandemic,” SSRN Electronic Journal, 2020, doi: 10.2139/ssrn.3746111.

[9] S. Salarvand, M.-S. Mousavi, and M. Rahimi, “Communication and cooperation challenges in the online classroom in the COVID-19 era: A qualitative study,” BMC Medical Education, vol. 23, no. 1, Art. 369, 2023, doi: 10.1186/s12909-023-04480-z.

[10] E. Alcalde Peñalver and J. García Laborda, “Online learning during the COVID-19 pandemic: How has this new situation affected students’ oral communication skills?,” Journal of Language and Education, vol. 7, no. 4, pp. 30–41, Dec. 2021, doi: 10.17323/jle.2021.11940.

[11] M. Zamiri and A. Esmaeili, “Strategies, methods, and supports for developing skills within learning communities: A systematic review of the literature,” Administrative Sciences, vol. 14, no. 9, art. 231, 2024, doi: 10.3390/admsci14090231.

[12] A. Lipsey, J. Irwin, and S. Coronel, Breaking Down Barriers to Career Development, Future Skills Centre & Blueprint, Ottawa, ON, Canada, Nov. 2021. [Online]. Available: https://fsc-ccf.ca/wp-content/uploads/2021/11/FSC-RCP-Barriers-EN.pdf. Accessed: Jan. 9, 2026.

[13] Statistics Canada, Education Indicators in Canada: An International Perspective 2017 — Chapter E: Participation in Formal and/or Non-Formal Education, Catalogue no. 81-604-X, Ottawa, ON, Canada, Feb. 14, 2018. [Online]. Available: https://www150.statcan.gc.ca/n1/pub/81-604-x/2017001/ch/che-eng.htm. Accessed: Jan. 9, 2026.

[14] H. J. Passow, "What competencies should engineering programs emphasize? A meta-analysis of practitioners' opinions informs curricular design," in Proc. 3rd Int. CDIO Conf., MIT, Cambridge, MA, USA, Jun. 11–14, 2007. [Online]. Available: https://www.cdio.org/files/document/file/T2A1Passow.pdf. Accessed: Jan. 10, 2026.

[15] L. J. Shuman, M. Besterfield-Sacre, and J. McGourty, "The ABET 'professional skills'—Can they be taught? Can they be assessed?," Journal of Engineering Education, vol. 94, no. 1, pp. 41–55, 2005.

[16] S. A. Male, M. B. Bush, and E. S. Chapman, "An Australian study of generic competencies required by engineers," European Journal of Engineering Education, vol. 36, no. 2, pp. 151–163, 2011, doi: 10.1080/03043797.2011.569703.

[17] Orai, AI-powered public speaking coach. [Online]. Available: https://orai.com/. Accessed: Jan. 10, 2026.

[18] Yoodli, AI speech coach for public speaking & interviews. [Online]. Available: https://yoodli.ai/. Accessed: Jan. 10, 2026.

[19] Google, Interview Warmup: Practice answering interview questions. [Online]. Available: https://grow.google/interview-warmup/. Accessed: Jan. 10, 2026.

[20] S. Chen, J. Zhou, X. Xu, X. Yang, L. Guo, and Y.-C. Chen, "PresentCoach: Dual-Agent presentation coaching through exemplars and interactive feedback," arXiv, Nov. 19, 2025, arXiv:2511.15253 [cs.HC]. [Online]. Available: https://arxiv.org/pdf/2511.15253.pdf. Accessed: Jan. 10, 2026.

[21] L. Alexandra, "Nonverbal cue analysis in video interviews," ResearchGate Preprint, Jun. 2025. [Online]. Available: https://www.researchgate.net/publication/392404491_Nonverbal_Cue_Analysis_in_Video_Interviews. Accessed: Jan. 10, 2026.

[22] M. E. Hoque, M. Courgeon, J.-C. Martin, B. Mutlu, and R. W. Picard, "MACH: My automated conversation coach," in Proc. ACM Int. Joint Conf. Pervasive and Ubiquitous Computing (UbiComp '13), Zurich, Switzerland, 2013, pp. 1–10. [Online]. Available: https://alumni.media.mit.edu/~mehoque/Publications/13.Hoque-etal-MACH-UbiComp.pdf. Accessed: Jan. 10, 2026.

[23] I. de Kok, "Presentation trainer: Your public speaking multimodal coach," ResearchGate, 2015. [Online]. Available: https://www.researchgate.net/publication/283703737_Presentation_Trainer_your_Public_Speaking_Multimodal_Coach. Accessed: Jan. 10, 2026.

[24] B. W. Schuller, S. Steidl, A. Batliner, F. Burkhardt, L. Devillers, C. Müller, and S. S. Narayanan, "Paralinguistics in speech and language – State-of-the-art and the challenge," Computer Speech & Language, vol. 27, no. 1, pp. 4–39, Jan. 2013, doi: 10.1016/j.csl.2012.02.005.

[25] I. Johnson and A. Smith, “LLMs as tools for evaluating textual coherence: A comparative analysis,” ResearchGate Preprint, 2025. [Online]. Available: https://www.researchgate.net/publication/385938814_LLMs_as_Tools_for_Evaluating_Textual_Coherence_A_Comparative_Analysis. Accessed: Jan. 11, 2026.’

[26] C. Graham and N. Roll, “Evaluating OpenAI’s Whisper ASR: Performance analysis across diverse accents and speaker traits,” JASA express letters, vol. 4, no. 2, Feb. 2024, doi: https://doi.org/10.1121/10.0024876.

[27] A. Radford, J. Kim, T. Xu, G. Brockman, C. Mcleavey, and I. Sutskever, “Robust Speech Recognition via Large-Scale Weak Supervision,” Dec. 2022. Available: https://cdn.openai.com/papers/whisper.pdf

[28] Gabriela M. Stegmann et al., “Repeatability of Commonly Used Speech and Language Features for Clinical Applications,” Digital Biomarkers, vol. 4, no. 3, pp. 109–122, Dec. 2020, doi: https://doi.org/10.1159/000511671.

[29] F. Eyben et al., “The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) for Voice Research and Affective Computing,” IEEE Transactions on Affective Computing, vol. 7, no. 2, pp. 190–202, Apr. 2016, doi: https://doi.org/10.1109/taffc.2015.2457417.

[30] Jin Hyun Cheong, T. Xie, S. Byrne, and L. J. Chang, “Py-Feat: Python Facial Expression Analysis Toolbox,” Affective Science, Aug. 2023, doi: https://doi.org/10.1007/s42761-023-00191-4.

[31] T. K. Koo and M. Y. Li, “A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research,” Journal of Chiropractic Medicine, vol. 15, no. 2, pp. 155–163, Jun. 2016, doi: https://doi.org/10.1016/j.jcm.2016.02.012.

Appendices

Appendix A: Test Procedures and Setup Details

This appendix summarizes all planned system tests, including their objectives, required inputs/tools, expected outputs, and strict pass/fail criteria used to validate end-to-end functionality, performance, privacy compliance, and scoring accuracy of the multimodal feedback tool.

Test Category	Objective	Inputs / Tools / Setup	Output Data	Pass/Fail Criteria
End-to-End Pipeline Test	Confirm full pipeline functionality: MP4 → multimodal extraction → features → JSON report	Inputs: Participant MP4 recordings (1–2 mins) Tools: - Smartphones or webcams. - Whisper (transcript generation) - OpenSMILE (audio feature extraction) - Py-Feat (visual feature extraction) Setup: Backend server and JSON viewer for inspection	Extracted audio, video, transcript streams; JSON report with ≥5 scoring categories and timestamps	- MP4 processed streams extracted - JSON report contains ≥5 scoring categories - Timestamped feedback aligns with observable events
Video Size & Duration Constraint Test	Ensure the system enforces the max size/duration	Inputs: MP4 files between 1–10 minutes, including boundary cases near 10 min/1GB Setup: upload valid and oversized files	Successful processing of valid files; rejection logs for oversized files	Accepts ≤10 min & ≤1GB; rejects larger videos
Runtime Performance Test	Confirm pipeline processes inputs within the required time	Inputs: MP4 files that are 1 GB Tools: Timestamp logging; Setup: Upload a sample MP4 file to measure execution time	Processing time per trial	Processes ≤10 min / 1GB video in <60 seconds

Test Category	Objective	Inputs / Tools / Setup	Output Data	Pass/Fail Criteria
Storage & Privacy Test	Verify that user files are deleted after the session	Tools: Storage/log monitoring Setup: Repeated uploads and check deletion time	Storage logs showing deletion	Ensure files are deleted after the report is generated. Confirm no persistent storage.
Usability & User Experience Test	Measure user perception of usefulness & accuracy	Setup: Participants receive a JSON report, and complete the Google Form rating ease, usefulness, accuracy, and likelihood of reuse	Survey results (numeric scores and comments)	≥70% rate usefulness ≥7/10 and majority indicate willingness to reuse
Human Ground-Truth Accuracy Test	Validate the correctness of feedback vs. human reviewers	Tools: "Ideal answer" checklist Setup: Human evaluators score videos using a rubric and an "ideal answer" checklist	Agreement % between tool output and human scoring	≥80% agreement across scored items
Filler-Word Accuracy Test	Validate detection of filler words	Inputs: Scripted videos with known filler-word counts	Detected vs. true counts	≥80% correct detection
Timestamp Accuracy Test	Ensure timestamps align with real events	Tools: Video playback, JSON viewer Setup: Side-by-side comparison of visible events vs JSON timestamps	Timestamp alignment	Timestamps match visible/audio events

Appendix B

This appendix presents a comparative evaluation of the language models considered for integration within the multimodal feedback pipeline. The comparison includes performance metrics such as reasoning benchmarks (MMLU-Pro, GPQA Diamond), inference speed, and approximate cost per call. These results informed the selection of the backend model by balancing performance, responsiveness, and computational efficiency for real-time educational feedback applications.

Table B1. Comparative performance of candidate large language models across benchmark accuracy, inference speed, and estimated cost per call.

	Speed (output tokens/second)	MMLU-Pro	Humanity's Last Exam	GPQA Diamond	Pricing (avg USD per 1M tokens)	Price on avg per call
Deepseek R1 0528 Qwen3 8B (free)	**64**	**74**	**5.6**	**61.1**	**0.07**	0.00007
Llama 3.3 70B Instruct (free, not opensource)	122	71	4	50	0.6	0.0006
DeepSeek: R1 0528 (free)	**28**	**85**	**17.7**	**81**	**0.1**	0.0001
Mistral: Devstral Small (free)	**138**	**63**	**4**	**43**	**0.15**	0.00015
DeepSeek v3 (free)	32	82	5.2	66	0.5	0.0005
Llama 4 Maverick (free, not opensource)	169	81	4.8	67	0.4	0.0004
Ministral 3b	**252**	**34**	**5.5**	**26**	**0.04**	0.00004
Ministral 8b	**191**	**39**	**4.9**	**28**	**0.1**	0.0001

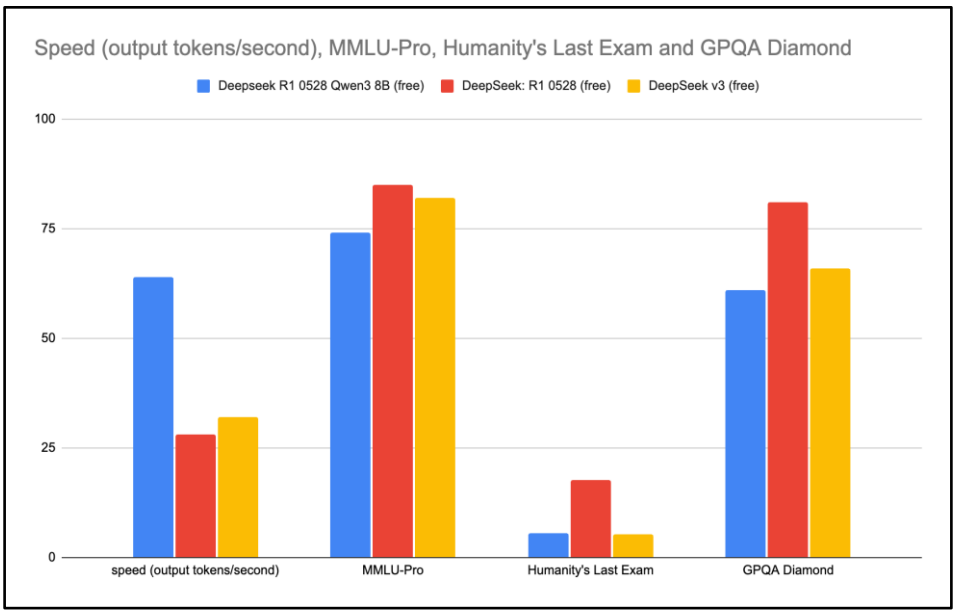

Figure B1. Performance comparison of candidate LLMs across accuracy benchmarks and inference efficiency, highlighting trade-offs between model capability and deployment cost.