



ASEAN Artificial Intelligence Journal

Journal homepage:
<https://karyailham.com.my/index.php/aiij/index>
 ISSN: 3083-9971



Fluency.my: An Artificial Intelligence-Driven Multimodal Presentation Coaching Tool for Malaysian University English Test Preparation

Ziyan Nifail¹, Muhammad Irsyad Abdullah^{1,*}, Asmaa Mahfoud Hezam Al-Hakimi¹

¹ Faculty of Information Sciences and Engineering, Management and Science University, Section 13, 40100 Shah Alam, Selangor, Malaysia

ARTICLE INFO

Article history:

Received 26 July 2026
 Received in revised form 30 September 2026
 Accepted 30 September 2026
 Available online 9 October 2026

Keywords:

Malaysian University English Test;
 automated speaking assessment;
 multimodal learning analytics; foreign
 language speaking anxiety; formative
 feedback

ABSTRACT

Malaysian learners preparing for the Malaysian University English Test face restricted access to individualised speaking rehearsal and feedback, alongside anxiety that can discourage oral practice. This study evaluated whether Fluency.my, a locally oriented multimodal speaking coach, provided a usable practice environment, reduced immediate self-reported speaking anxiety, and supported automated analysis of learner recordings. [MY INFERENCE] Design science research and multimodal learning analytics framed the platform as an educational artefact whose value depended on the joint operation of linguistic, vocal, and visual feedback. A mixed-methods pilot evaluation was conducted with 10 students at Management and Science University in Malaysia after each participant completed a platform session. The analysis descriptively examined System Usability Scale responses, adapted Foreign Language Classroom Anxiety Scale scores collected before and after practice, and speech-recognition and computer-vision performance metrics. The participants' mean System Usability Scale score was 74.5. The mean adapted Foreign Language Classroom Anxiety Scale score decreased from 24.2 before practice to 19.1 afterwards, corresponding to the reported 21.1% immediate reduction. The speech-recognition pipeline recorded a word error rate of 11.4%, while face detection reached 94.2%. [MY INFERENCE] The study's originality lies in integrating examination-aligned language coaching with speech, voice, and non-verbal analysis for Malaysian learners. [MY INFERENCE] Educators should use the platform as a supervised rehearsal aid while retaining human judgement for consequential speaking assessment.

1. Introduction

For Malaysian university students, preparation for the Malaysian University English Test speaking component requires repeated oral practice, criterion-aligned feedback, and opportunities to rehearse under manageable evaluative pressure. These conditions are difficult to reproduce when individual feedback is scarce or available only during formal teaching. Speaking anxiety compounds the problem because fear of negative evaluation can suppress voluntary practice before the assessment itself. Foreign language classroom anxiety is established as a distinct affective construct

* Corresponding author.
 E-mail address: irsyad@msu.edu.my

[1], and earlier reviews describe its disruptive relationship with language learning and performance [2].

Technology-mediated rehearsal can create a lower-exposure environment in which learners repeat speaking tasks and receive feedback without an immediate human audience. The evidence is not uniform. High-immersion virtual reality reduced anxiety relative to videoconferencing in one repeated-practice study [3], whereas artificial intelligence-supported instruction improved speaking outcomes without changing anxiety in another study [4]. A randomised high-immersion virtual-reality study also found that lower anxiety within the virtual setting did not transfer reliably to real-life speaking [5]. Task-specific analysis further shows that anxiety change can differ across pair work, presentations, debates, and storytelling [6]. Systematic-review and meta-analytic evidence therefore supports a conditional account in which technology can reduce anxiety, but its effect depends on psychological safety, feedback, task structure, and social interaction [7,8].

Automated speaking support also presents a measurement problem. Multimodal learning analytics combines evidence from more than one channel to represent complex performance [9], while contemporary language systems can add transcript-based scoring and generated formative comments. Research on automated scoring shows that a language model can be constrained by structured criteria [10], but recognition accuracy remains sensitive to accent, first-language phonology, proficiency, speaker traits, and speech style [11]. Domain-specific fine-tuning can improve recognition for non-native speech, although performance changes with the composition and preparation of the training corpus [12].

The unresolved issue is therefore not simply whether a digital tool can produce feedback, but whether exam-aligned linguistic scoring, vocal analysis, non-verbal analysis, and affective support can operate together in one localised practice cycle. Existing evidence does not establish that a favourable user experience, lower immediate anxiety, and acceptable transcription accuracy will co-occur for Malaysian learners preparing for the same assessment. This integration is theoretically important because affective benefit should depend on both the learning environment and the dependability of the automated interpretation on which feedback is based.

The study consequently designed and evaluated Fluency.my, an artificial intelligence-driven multimodal platform for Malaysian University English Test speaking preparation. It examined whether the prototype could complete an examination-aligned coaching cycle, whether participants found it usable, whether self-reported speaking anxiety changed immediately after one session, and how the speech-recognition and computer-vision components performed. The contribution was a localised artefact and a pilot evaluation that connected system feasibility, user experience, affective response, and recognition performance within the same study.

2. Methodology

2.1 Research Design

The study adopted design science research within a mixed-methods pilot framework. Design science research was appropriate because the work centred on constructing and evaluating an educational technology artefact [13]. Development proceeded through iterative software sprints in which platform functions and analysis outputs were reviewed against system development, model implementation, and testing objectives.

2.2 Participants, Setting, and Procedure

The pilot involved 10 student participants at Management and Science University in Malaysia. Participants were grouped in the source data as above-average, average, or below-average proficiency. Each participant completed a platform session, responded to the System Usability Scale after use, and completed the adapted Foreign Language Classroom Anxiety Scale immediately before and after the session. The evaluation also recorded performance metrics from the speech-recognition and computer-vision pipeline.

2.3 System Architecture

Fluency.my was implemented as a full-stack web application. The frontend used Next.js 15, React 19, TypeScript, and Tailwind CSS, while the backend used FastAPI and Uvicorn. Supabase PostgreSQL provided the database layer, row-level security, and private video storage. Table 1 summarises the technologies used in the platform.

Table 1
Technology stack

Layer	Technology
Frontend	Next.js 15 (App Router), React 19, TypeScript, Tailwind CSS
Backend	FastAPI and Uvicorn (Python asynchronous processing)
Database and storage	Supabase PostgreSQL with row-level security; Supabase Storage
Speech-to-text	Groq whisper-large-v3-turbo; local OpenAI Whisper fallback
Language-model coaching	Groq llama-3.3-70b-versatile
Computer vision	MediaPipe 0.10.x Face Mesh and Pose
Voice analysis	librosa pitch estimation and root-mean-square energy
Sentiment	Valence Aware Dictionary and Sentiment Reasoner

2.4 Artificial Intelligence Analysis Pipeline

The analysis pipeline was triggered after video upload and executed asynchronously. MediaPipe vision analysis, librosa voice analysis, and Whisper transcription ran in parallel. The completed transcript was then analysed for speaking pace, filler density, lexical diversity, discourse markers, sentence-length variation, and sentiment. Figure 1 shows the sequence from upload through feature extraction, scoring, feedback generation, and storage.

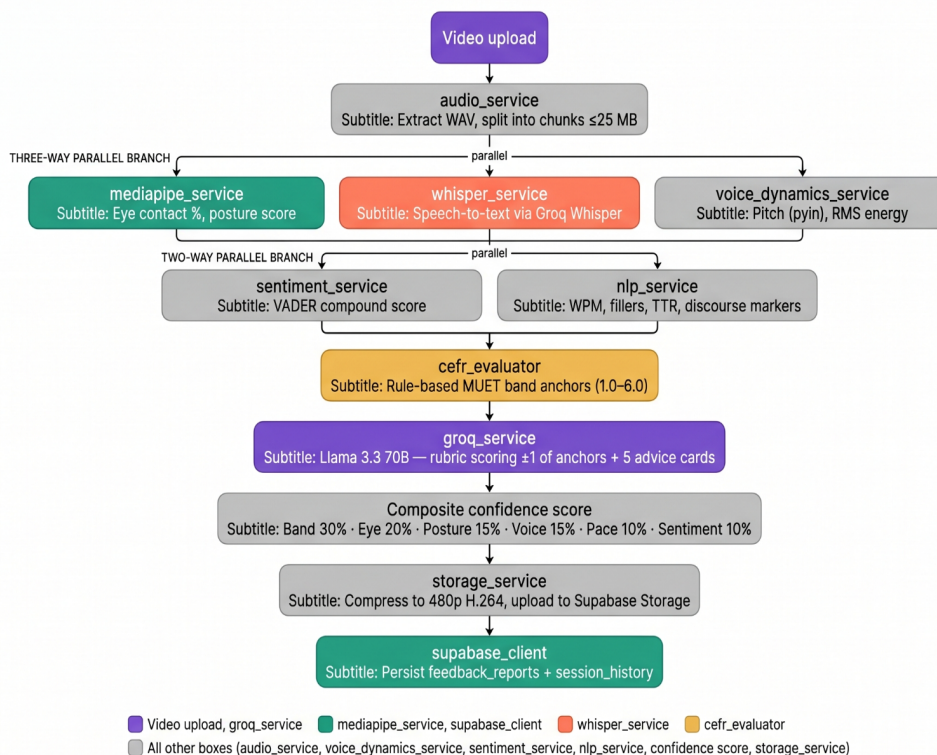


Fig. 1. Fluency.my artificial intelligence analysis pipeline

A two-layer scoring process combined rule-based anchors with language-model evaluation. The rule layer generated criterion anchors from linguistic and delivery measures. The language model received the anchors and transcript and was constrained to score within the tolerance implemented in the source system. The platform then generated ranked advice and a composite confidence score from the band, eye-contact, posture, voice, pace, sentiment, and filler-density measures.

2.5 Coaching and Feedback Functions

During guided sessions, the platform monitored gaze, posture, filler words, and speaking pace. After a session, it presented a Malaysian University English Test band estimate, criterion-level feedback, a transcript with highlighted language features, pace information, and prioritised advice. Educator functions supported task assignment, review of submissions, video playback, and documented human correction of an automated band estimate.

2.6 Measures and Analysis

Usability was measured with the 10-item System Usability Scale [14]. Speaking anxiety was measured with an adapted 8-item Foreign Language Classroom Anxiety Scale administered before and after the session. The original scale is established in language-anxiety research [1], while validation work on a short form shows that item reduction requires separate psychometric examination [15]. Speech-recognition performance was evaluated with word error rate, and visual performance was evaluated with face- and pose-detection rates. The study reported descriptive values and participant-level changes; it did not conduct an inferential test of the anxiety outcome.

3. Results and Discussion

3.1 Artificial Intelligence Pipeline Performance

The speech-recognition component recorded a word error rate of 11.4% across the test recordings. Face detection reached 94.2%, pose detection reached 91.7%, and the average posture score was 74.2 out of 100. Repeated submission of identical inputs produced a maximum band deviation of ± 0.3 , and the end-to-end pipeline processed a 2-minute video in 18.4 seconds. Table 2 presents the complete pipeline metrics.

Table 2
Artificial intelligence pipeline performance

Metric	Result
Word error rate	11.4%
Face detection rate	94.2%
Pose detection rate	91.7%
Average posture score	74.2 / 100
Band-score consistency, maximum deviation	± 0.3 bands
End-to-end pipeline time for a 2-minute video	18.4 seconds

3.2 System Usability

The overall mean System Usability Scale score was 74.5. Participant scores ranged from 52.5 to 90.0 across the proficiency groups recorded in the source data. Table 3 reports each participant's score and the classification used in the pilot.

Table 3
System usability scale scores by participant

Code	Group	Score (/100)	Usability grade
P01	Above Average	87.5	Excellent
P02	Above Average	82.5	Excellent
P03	Above Average	90.0	Excellent
P04	Above Average	80.0	Excellent
P05	Above Average	85.0	Excellent
P06	Average	72.5	Above Average
P07	Average	67.5	Average
P08	Average	70.0	Above Average
P09	Below Average	57.5	Below Average
P10	Below Average	52.5	Below Average
Overall	—	74.5	Above Average

3.3 Speaking Anxiety

All 10 participants recorded lower self-reported anxiety after the session. The mean adapted Foreign Language Classroom Anxiety Scale score moved from 24.2 before practice to 19.1 afterwards. The source Results reported this change as a 21.1% mean reduction. Table 4 gives the participant-level results, and Figure 2 displays the before-and-after pattern. No inferential test was reported.

Table 4
Adapted foreign language classroom anxiety scale results

Code	Group	Pre-test (/40)	Post-test (/40)	Change	Reduction
P01	Above Average	15	9	-6	40.0%
P02	Above Average	18	11	-7	38.9%
P03	Above Average	14	8	-6	42.9%
P04	Above Average	20	14	-6	30.0%
P05	Above Average	21	15	-6	28.6%
P06	Average	26	22	-4	15.4%
P07	Average	28	22	-6	21.4%
P08	Average	28	24	-4	14.3%
P09	Below Average	34	31	-3	8.8%
P10	Below Average	38	35	-3	7.9%
Overall Average	—	24.2	19.1	-5.1	21.1%

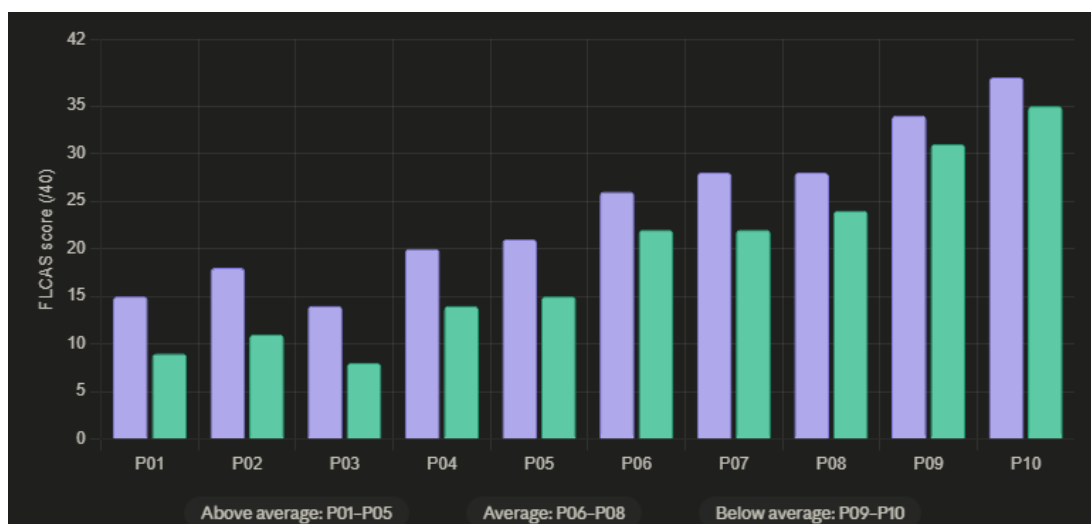


Fig. 2. Adapted foreign language classroom anxiety scale scores before and after the session

3.4 Discussion

The three outcomes address different parts of feasibility. The usability score describes the participants' experience of the interface and workflow, the anxiety scores describe immediate self-report change, and the word error rate describes transcription performance under the pilot conditions. Together, they indicate that the prototype completed a usable coaching cycle and that

the affective outcome moved in the intended direction. They do not establish an intervention effect, population-level usability, or equivalence between automated and human assessment.

The usability result was broadly consistent with recent educational-technology evidence. A systematic review reported an overall System Usability Scale mean of 64.5 with a standard deviation of 14.61 [16], while a mobile language-learning task recommender reported a mean of 72.0 [17]. A study of ChatGPT for formal English learning reported its usability result on a different denominator, preventing direct comparison without clarification [18]. [MY INFERENCE] Fluency.my's broader workflow may impose different demands because users must record speech, await multimodal processing, interpret band-aligned feedback, and navigate several forms of analysis. The pilot size also means that the mean should be treated as a feasibility estimate rather than a stable population value.

The immediate decrease in adapted anxiety scores followed the direction reported in parts of the digital speaking-practice literature. Virtual-reality practice reduced anxiety relative to videoconferencing in one repeated-measures study [3], and the meta-analysis reported an overall Hedges' g of -0.522 and a sensitivity estimate of -0.441 [8]. Other studies established important boundaries: one artificial intelligence-supported intervention improved speaking without changing anxiety [4], reduced anxiety in a virtual setting did not transfer reliably to real-life speaking [5], and anxiety change varied by task type [6]. Fluency.my may have lowered immediate evaluative pressure through private practice and automated feedback, but the current design cannot determine whether that change persists or transfers to an examiner-facing task.

The word error rate was comparable in magnitude with selected Whisper results under some conditions but cannot be treated as a universal accuracy estimate. Graham and Roll reported substantial variation by first-language phonology, proficiency, speaker traits, and speech style [11]. Jain and colleagues showed that domain-specific fine-tuning and corpus composition changed recognition performance across native and non-native child-speech datasets [12]. The next system evaluation should therefore report error separately by proficiency, Malaysian English variety, code-switching, and recording condition to determine whether the aggregate value conceals unequal performance.

The design also limits causal interpretation. A one-group pre-test–post-test design remains vulnerable to testing and other internal-validity threats [19], and pilot sample requirements depend on the estimation goal of the subsequent study [20]. The current evidence supports feasibility rather than effectiveness. A controlled, longitudinal comparison with blinded human ratings is required before the platform can be claimed to improve Malaysian University English Test speaking performance or produce durable anxiety reduction.

4. Conclusions

This study examined whether a locally designed digital speaking coach could offer Malaysian learners a workable form of Malaysian University English Test preparation while easing the practical and emotional barriers that arise when personalised rehearsal and feedback are difficult to obtain. The pilot showed that learners were able to use the platform, that self-reported speaking apprehension moved downward immediately after practice, and that the integrated speech and visual-analysis pipeline completed the coaching process as intended. These findings establish feasibility, not effectiveness: they show that the prototype can support a complete practice cycle, but they do not demonstrate durable anxiety change, better examination performance, equitable recognition across learner groups, or agreement with expert scoring. For theory, the study positions affective benefit as the product of an interaction among the learning setting, the feedback encounter,

and the dependability of automated interpretation, rather than as an inherent property of educational artificial intelligence. Educators and institutional decision-makers should deploy the system as a supervised rehearsal aid, retain human judgement for consequential assessment, and monitor whether lower-proficiency and locally accented users receive feedback of comparable quality. The priority is a multi-wave comparison against usual Malaysian University English Test preparation, with blinded examiner judgements, anxiety assessed under examination-like conditions, and transcription errors analysed separately for proficiency and Malaysian English variety.

Acknowledgement

The authors acknowledge the Faculty of Information Sciences and Engineering at Management and Science University for the institutional support provided to the project and thank the student participants who completed the pilot evaluation.

References

- [1] Horwitz, Elaine K., Michael B. Horwitz, and Joann Cope. "Foreign language classroom anxiety." *The Modern language journal* 70, no. 2 (1986): 125-132. <https://doi.org/10.1111/j.1540-4781.1986.tb05256.x>
- [2] Scovel, Thomas. "The effect of affect on foreign language learning: A review of the anxiety research." *Language learning* 28, no. 1 (1978): 129-142. <https://doi.org/10.1111/j.1467-1770.1978.tb00309.x>
- [3] Kaplan-Rakowski, Regina, and Alice Gruber. "The impact of high-immersion virtual reality on foreign language anxiety." *Smart Learning Environments* 10, no. 1 (2023): 46. <https://doi.org/10.1186/s40561-023-00263-9>
- [4] Qiao, Hongliang, and Aruna Zhao. "Artificial intelligence-based language learning: illuminating the impact on speaking skills and self-regulation in Chinese EFL context." *Frontiers in psychology* 14 (2023): 1255594. <https://doi.org/10.3389/fpsyg.2023.1255594>
- [5] Ding, Miaomiao. "The impact of high-immersion virtual reality on EFL learners' foreign language speaking anxiety: A mixed-method approach." *ReCALL* 36, no. 3 (2024): 287-305. <https://doi.org/10.1017/S0958344024000156>
- [6] Huang, Yanbin, and Phadet Kakham. "The impacts of AI conversational agents on EFL learners' oral proficiency and foreign language speaking anxiety." In *Frontiers in Education*, vol. 11, p. 1799269. Frontiers Media SA, 2026. <https://doi.org/10.3389/feduc.2026.1799269>
- [7] Huang, Lujie, and Ming Liu. "The impact of technology on foreign language anxiety: A systematic review of empirical studies from 2004 to 2024." *Humanities and Social Sciences Communications* 12, no. 1 (2025): 1665. <https://doi.org/10.1057/s41599-025-05921-6>
- [8] Kayhan, Selvanur. "The effect of intervention studies on foreign language speaking anxiety: A meta-analysis study." *Journal of Pedagogical Research* 9, no. 3 (2025): 62-83. <https://doi.org/10.3390/JPR.202533156>
- [9] Blikstein, Paulo, and Marcelo Worsley. "Multimodal learning analytics and education data mining: Using computational technologies to measure complex learning tasks." *Journal of learning analytics* 3, no. 2 (2016): 220-238. <https://doi.org/10.18608/jla.2016.32.11>
- [10] Mizumoto, Atsushi, and Masaki Eguchi. "Exploring the potential of using an AI language model for automated essay scoring." *Research Methods in Applied Linguistics* 2, no. 2 (2023): 100050. <https://doi.org/10.1016/j.rmal.2023.100050>
- [11] Graham, Calbert, and Nathan Roll. "Evaluating OpenAI's Whisper ASR: Performance analysis across diverse accents and speaker traits." *JASA Express Letters* 4, no. 2 (2024) <https://doi.org/10.1121/10.0024876>
- [12] Rishabh Jain, Andrei Barcovschi, Mariam Yahayah Yiwere, Peter Corcoran, and Horia Cucu. "Exploring Native and Non-Native English Child Speech Recognition with Whisper." *IEEE Access* 12 (2024): 41601–41610. <https://doi.org/10.1109/ACCESS.2024.3378738>
- [13] Hevner, Alan R., Salvatore T. March, Jinsoo Park, and Sudha Ram. "Design science in information systems research1." *MIS quarterly* 28, no. 1 (2004): 75-106. <https://doi.org/10.2307/25148625>
- [14] John Brooke. "SUS: A 'Quick and Dirty' Usability Scale." In *Usability Evaluation in Industry*, 189–194. 1996. <https://doi.org/10.1201/9781498710411-35>
- [15] Elouise Botes, Lindie van der Westhuizen, Jean-Marc Dewaele, Peter MacIntyre, and Samuel Greiff. "Validating the Short-Form Foreign Language Classroom Anxiety Scale." *Applied Linguistics* 43, no. 5 (2022): 1006–1033. <https://doi.org/10.1093/applin/amac018>

- [16] Vlachogianni, Prokopia, and Nikolaos Tselios. "Perceived usability evaluation of educational technology using the System Usability Scale (SUS): A systematic review." *Journal of Research on Technology in Education* 54, no. 3 (2022): 392-409. <https://doi.org/10.1080/15391523.2020.1867938>
- [17] Yudhy Setyo Purwanto, Rahmat Gernowo, and Dinar Mutiara Kusumo Nugraheni. "ACTE: A Pilot Feasibility Evaluation of a Mastery-Aware Task Recommender for Mobile Language Learning in Real-World Contexts." *Journal of Information Systems and Informatics* 8, no. 2 (2026): 1640–1671. <https://doi.org/10.63158/journalisi.v8i2.1554>
- [18] Sarang Shaikh, Sule Yildirim Yayilgan, Blanka Klimova, and Marcel Pikhart. "Assessing the Usability of ChatGPT for Formal English Language Learning." *European Journal of Investigation in Health, Psychology and Education* 13, no. 9 (2023): 1937–1960. <https://doi.org/10.3390/ejihpe13090140>
- [19] Bierer, S. Beth, Gary Beck Dallaghan, Nicole J. Borges, Sam Brondfield, Cha Chi Fung, Kathryn N. Huggett, Cayla R. Teal, Satid Thammasitboon, and Colleen Y. Colbert. "Moving beyond simplistic research design in health professions education: What a one-group pretest-posttest design will not prove." *MedEdPORTAL* 21 (2025): 11527. https://doi.org/10.15766/mep_2374-8265.11527
- [20] Whitehead, Amy L., Steven A. Julious, Cindy L. Cooper, and Michael J. Campbell. "Estimating the sample size for a pilot randomised trial to minimise the overall trial sample size for the external pilot and main trial for a continuous outcome variable." *Statistical methods in medical research* 25, no. 3 (2016): 1057-1073. <https://doi.org/10.1177/0962280215588241>