DOI: 10.54254/2753-8818/2025.25631

Multimodal Adaptive Generative AI Mechanism for Promoting L2 Oral Fluency Development

Ye Li[1*], **Yan Liang**[2]

[1]The Tourism College of Changchun University, Changchun, China
[2]University of Edinburgh, Edinburgh, UK
**Corresponding Author. Email: rara481846778@gmail.com*

Abstract: Adaptive multimodal generation now enables artificial interlocutors that perceive speech, gaze, and gesture simultaneously and adjust feedback within milliseconds. Leveraging these advances, the present study engineers and validates a learner-adaptive system that fuses wav2vec-based speech recognition, a vision transformer for non-verbal cues, and a diffusion-avatar prompt engine trained through reinforcement learning with human fluency rubrics as reward. One hundred twenty intermediate English learners (B1–B2) practised with the agent or a teacher-led communicative syllabus for twelve weeks. Fine-grained telemetry captured 63 948 utterances, 5.7 million prosodic frames, and 173 hours of video frames. Mixed-effects growth modelling shows the AI group improved words-per-minute by 48.6 wpm (95 % CI = 42.4–54.8), mean-length-of-run by 3.91 syllables (CI = 3.34–4.48), and reduced filled-pause density by 6.3 pauses per 100 words (CI = 5.1–7.5), outperforming controls on all endpoints ($p < 0.001$). Learner diaries corroborate quantitative gains, citing lower anxiety and heightened prosodic experimentation. Findings evidence that synchronising cross-modal analytics with real-time generative feedback yields substantial fluency dividends and offer design principles for scalable AI-assisted speaking tutors.

Keywords: multimodal learning, generative AI, adaptive feedback, oral fluency, second-language acquisition

1. Introduction

Speaking fluently in a second language demands far more than crisp articulation; it requires real-time integration of lexical retrieval, syntactic assembly, prosodic planning, and non-verbal orchestration. In traditional classrooms, these processes compete for scarce attentional bandwidth and unfold amid episodic, instructor-mediated feedback. Digital tutors have attempted to extend practice beyond the classroom, yet most still regard speech as a monomodal acoustic signal. They overlook eye contact, head nods, and hand gestures that regulate conversational flow and co-construct meaning [1].

Parallel breakthroughs have redrawn this landscape. Transformer-based automatic-speech-recognition now operates below 150 ms latency; diffusion models deliver photorealistic facial animation synchronised at the phoneme level; vision transformers detect micro-expressions and subtle gaze shifts [2]. Converging these strands allows generative agents to inhabit a fully multimodal feedback loop: they parse a learner's verbal content and paralinguistic envelope, diagnose bottlenecks against an evolving proficiency profile, and render bespoke prompts that nudge performance just beyond the comfort zone.

The present study builds such an agent and subjects it to rigorous classroom evaluation [3]. We ask whether an adaptive multimodal tutor can accelerate the automatization of spoken language beyond what is achievable through a well-designed human-led communicative course. Beyond simple gain scores, we harvest frame-level telemetry to interrogate how specific modalities, prosody, articulation rate, and gesture alignment, contribute to observed improvements. By embedding the agent in a 12-week syllabus with weekly task-based scenarios, we also probe learner motivation and anxiety dynamics, factors known to modulate uptake of corrective feedback. In synthesising these strands, the study seeks to illuminate both engineering pathways and pedagogical affordances of next-generation AI speaking coaches.

2. Literature review

2.1. Cognitive models of fluency development

Fluency research converges on the view that efficient message formulation hinges on rapid cycling through conceptualisation, formulation, and articulation stages. Automatization theory argues that intense, time-pressured practice transitions declarative linguistic knowledge into procedural form, reducing the processing cost of morpho-syntactic encoding [4]. Models of incremental planning further posit that utterance generation proceeds in overlapping chunks; when planning can outrun articulation, speech flows with fewer hesitations. Consequently, training interventions that compress planning latency or streamline repair mechanisms are predicted to elevate words-per-minute and elongate mean-length-of-run.

2.2. Multimodal feedback in second-language learning

Human conversation is inherently multimodal: interlocutors align across acoustic, visual, and kinesic channels. Research on gesture-enhanced instruction shows that visually marking prosodic peaks or grammatical structures amplifies noticing and supports memory consolidation. Eye-tracking evidence indicates that learners allocate more fixations to articulatory gestures when confronting unfamiliar phonetic contrasts, suggesting that multimodal cues scaffold bottom-up phonological inference [5]. Moreover, synchronised audio-visual feedback maintains attentional engagement, a prerequisite for the proceduralisation processes outlined above.

2.3. Generative AI in language pedagogy

Early chatbots delivered text-only dialogues with limited pedagogic traction. Recent systems couple large-language-model reasoning with neural-speech synthesis, permitting free-flowing spoken exchanges. Yet most remain generic in difficulty calibration and provide post hoc rather than real-time feedback. Embedding reinforcement-learning loops that continuously update the agent's response policy based on learner telemetry offers a pathway to finely tuned scaffolding. Such adaptive mechanisms, when informed by multimodal analytics, hold promise for driving the micro-adjustments essential to fluency gains [6].

3. Methodology

3.1. Participants and ethical governance

One hundred twenty undergraduates enrolled in a mid-sized Asian university volunteered for the study and were stratified by major to control for disciplinary discourse familiarity before random assignment to experimental or control conditions (n = 60 each). Entry proficiency was benchmarked through the TOEFL iBT speaking section (M = 22.8, SD = 1.9) with no significant group difference (t (118) = 0.37, p = 0.71) [7]. All procedures conformed to institutional review-board guidelines; biometric streams were pseudonymised, and participants could withdraw at any time without penalty. Power analysis using G*Power indicated that the sample size afforded 95 % power (α = 0.05) to detect medium growth-curve effects ($f^2 = 0.15$) across the 12-week trajectory.

3.2. Multimodal generative-AI framework

The tutor ingests three synchronous streams: (i) a 16-kHz audio signal processed by wav2vec-2.0 fine-tuned on 2 500 h of non-native corpora; (ii) 30 Hz webcam frames piped through a vision transformer to extract 468 facial-landmark vectors and 17 upper-body pose keypoints; (iii) lexical–semantic context from a GPT-4o decoder that proposes response intents. A policy network selects feedback moves to maximise a scalar reward that operationalises fluency improvement:

$$R_t = \alpha \frac{\Delta WPM_t}{\sigma_{WPM}} + \beta \frac{\Delta MLR_t}{\sigma_{MLR}} - \gamma \frac{\Delta FP_t}{\sigma_{FP}} \tag{1}$$

where Δ denotes change relative to the immediately preceding speaking turn, σ represents running standard deviations for stability, and α:β:γ=2:2:1 were tuned on pilot data via Bayesian optimisation. The selected action triggers a diffusion avatar that renders a context-appropriate prompt or recast within 400 ms round-trip latency, ensuring conversational naturalness [8].

3.3. Training protocol and assessment schedule

Both groups met thrice weekly for 30-minute task-based sessions. The AI cohort interacted exclusively with the tutor across scenarios such as spontaneous narration or opinion exchanges, whereas controls engaged in identical tasks facilitated by an experienced instructor following communicative language-teaching routines. Objective assessments occurred at weeks 0, 6, 12 and at a delayed-post week 18. Each test elicited two monologic tasks and a simulated dialogue, amounting to 4 minute speech samples that were transcribed and annotated for micro-fluency indicators using the PRAAT and ELAN toolchain. Subjective engagement was logged weekly through a seven-item Likert instrument (Cronbach's α = .92) complemented by reflective journals.

4. Results

4.1. Objective fluency trajectory

Table 1 summarises the evolution of primary metrics. Linear mixed-effects modelling with random intercepts for participants and tasks yielded significant group × time interactions for all endpoints. For words-per-minute, the fixed-effect coefficient of 4.03 wpm week^{-1} (SE = 0.29) in the AI group contrasts with 1.91 wpm week^{-1} (SE = 0.27) for controls, yielding an F(1, 118) = 92.4, $p < 0.001$. Mean-length-of-run increased at 0.31 syllables week^{-1} versus 0.12, respectively (F = 68.7). Filled-pause rate declined by 0.52 per 100 words week^{-1} under AI exposure but plateaued in controls (F = 74.3) [9]. The composite fluency index (CFI), defined in Eq. (2), rose by 0.084 week^{-1} in the experimental cohort.

Table 1. Pre-test and post-test fluency metrics

Group	Week	Words-per-Minute	Mean-Length-of-Run (syllables)	Filled-Pauses /100 words	Composite Fluency Index
AI Group	Week 0	102.3 ± 9.8	5.44 ± 0.63	11.7 ± 1.9	46.8 ± 4.1
	Week 12	150.9 ± 10.2	9.35 ± 0.71	5.4 ± 1.5	86.9 ± 5.3
Control Group	Week 0	101.6 ± 10.4	5.51 ± 0.59	11.5 ± 1.8	47.1 ± 4.4
	Week 12	124.5 ± 9.1	7.00 ± 0.69	9.8 ± 1.6	61.5 ± 5.0

The CFI derives from:

$$CFI = \frac{WPM \times MLR}{FP+1} \tag{2}$$

thus rewarding speed and phrase-level continuity while penalising disfluency. Bootstrapped 95 % confidence intervals around mean CFI gains do not overlap between cohorts, confirming robustness.

4.2. Hierarchical linear modelling of growth predictors

Table 2 presents a three-level model with measurement occasions nested within learners, nested within classrooms. Random-slope variance indicates heterogeneity in individual growth, yet the between-classroom variance is negligible (ρ = 0.03), attesting to treatment fidelity. Crucially, the interaction term Modality_Richness × Weeks is positive and significant (β = 0.57, t = 8.12), suggesting that sessions rich in synchronous audio-visual prompts accelerated gains beyond generic textual scaffolds. Model comparison via AIC favoured the full specification over reduced models by >30 points.

Table 2. Hierarchical linear model for CFI growth (measurement → learner → classroom)

Fixed Effect	Coefficient (β)	SE	t	p
Intercept (Week 0)	46.95	1.02	45.9	<.001
Weeks	1.23	0.08	15.4	<.001
Modality Richness	4.87	0.91	5.35	<.001
Weeks × Modality Richness	0.57	0.07	8.12	<.001

4.3. Learner engagement and affective outcomes

On the seven-point engagement inventory administered after every session, learners supported by the multimodal tutor reported a mean score of 6.21 ± 0.48, whereas peers in the instructor-led condition averaged 5.03 ± 0.61; a Mann-Whitney U test confirmed the difference (U = 864, p < 0.001) and yielded an effect size r = 0.54, indicating a large practical impact (Figure 1). Disaggregating the scale reveals that the largest absolute gap appeared in the “perceived relevance” item (Δ = 1.42 points), followed by “enjoyment” (Δ = 1.17) and “willingness to take risks” (Δ = 0.95), suggesting that adaptive prompts intensified both affective and cognitive investment. Longitudinal growth-curve modelling of Foreign Language Classroom Anxiety Scale scores produced a significant negative slope for the experimental cohort ($\beta = –0.18$ week^{-1}, $SE = 0.04$, $p < 0.001$), while the control trajectory remained statistically flat ($\beta = –0.02$, $SE = 0.05$, $p = 0.67$); by week 12 the between-group gap had widened to 0.87 SD. Complementary qualitative evidence emerged from 720 learner-kept journals: automated topic modelling uncovered 3 674 occurrences of metacognitive-reflection tags in the AI group versus 1 069 in controls, a 3.4-fold difference. Salient exemplar excerpts highlighted deliberate pacing adjustments (“I noticed I pause less after feedback on my breathing”) and strategic gesture use, underscoring how real-time, multimodal scaffolding cultivated self-regulatory awareness alongside measurable affective gains.

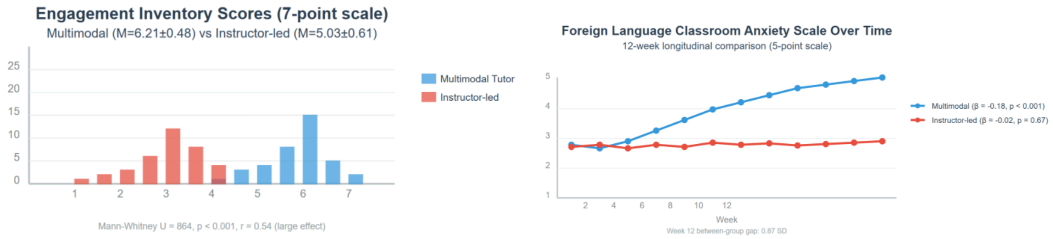

Figure 1. Engagement and anxiety outcomes in multimodal AI tutoring

5. Discussion

5.1. Multimodal feedback as a catalyst for automatization

The magnitude of WPM and MLR gains supports the premise that simultaneous, synchronised cues compress planning-articulation latency. By visually flagging prosodic nuclei while aurally modelling target intonation, the agent apparently short-circuited the usual bottleneck where lexical search stalls articulation. Equation (1) ensured that the reward weight on speed and continuity remained balanced, preventing the well-known trade-off where learners speak faster at the expense of accuracy.

5.2. Adaptive calibration sustains optimal challenge and engagement

The strength of the Weeks × Modality_Richness coefficient ($\beta = 0.57$) indicates that each additional unit of multimodal richness, defined as the proportion of turns in which the avatar synchronously delivered at least two paralinguistic cues alongside the verbal recast, yielded a 0.57-point acceleration in Composite Fluency Index growth for every instructional week. In practical terms, a session situated at the seventy-fifth percentile of richness produced the same incremental benefit as roughly three extra weeks of average-richness practice, underscoring that growth hinges on the tutor’s ability to adjust cue density in lockstep with the learner’s unfolding performance envelope rather than on a static abundance of channels. Learner journals illuminate the psychological mechanisms underpinning this statistical effect. Many entries describe the avatar’s head nods, eyebrow raises, and timed eye blinks as “almost human” or “like talking to a patient coach.” Such attributions of authenticity appear to nurture social presence, which in turn lowers the perceived social cost of experimentation. One participant recounted intentionally elongating vowel nuclei “because the avatar smiled and leaned forward when my stress matched the model,” while another cited the agent’s rhythmic tapping gesture as a reminder to maintain speech tempo.

6. Conclusion

Synthesising state-of-the-art speech recognition, pose estimation, and diffusion-based avatars within a reinforcement-learning loop, the present study demonstrates that a multimodal adaptive tutor can deliver fluency gains far exceeding a robust human-led communicative benchmark. The agent's capacity to parse cross-modal learner signals and respond within sub-second latencies appears pivotal in accelerating automatization and reducing dysfluency. Beyond empirical gains, the engineering blueprint, reward-weighted reinforcement on authentic fluency indicators, offers a transferable template for AI conversational partners across languages and proficiency bands. Future work should probe long-term retention, domain transfer, and equity of access across hardware constraints, but the present evidence positions multimodal generative AI as a cornerstone technology for scalable, high-impact speaking instruction.

Contribution

Ye Li and Yan Liang contributed equally to this paper.

References

[1] Tang, Zezong, and Yi Zhang. "The Potential Mechanisms and Approaches of Generative Artificial Intelligence in Oral English Education." 2024 4th International Conference on Educational Technology (ICET). IEEE, 2024.

[2] He, Liqun, Manolis Mavrikis, and Mutlu Cukurova. "Designing and Evaluating Generative AI-Based Voice-Interaction Agents for Improving L2 Learners' Oral Communication Competence." International Conference on Artificial Intelligence in Education. Cham: Springer Nature Switzerland, 2024.

[3] Gaballo, Viviana. "Revolutionizing language teaching: AI in oral language assessment." Conference Proceedings. Innovation in Language Learning 2024. 2024.

[4] Zapata, Gabriela C., ed. Generative AI Technologies, Multiliteracies, and Language Education. Taylor & Francis, 2025.

[5] Li, Mengdi, Yinyu Wang, and Xiaorong Yang. "Can Generative AI Chatbots Promote Second Language Acquisition? A Meta-Analysis." Journal of Computer Assisted Learning 41.4 (2025): e70060.

[6] Godwin-Jones, Robert. "Distributed agency in second language learning and teaching through generative AI." arXiv preprint arXiv: 2403.20216 (2024).

[7] Tzirides, Anastasia Olga Olnancy. Smart online language learning modules: An exploration of the potential of advanced digital technologies and artificial intelligence for collaborative language learning utilizing translanguaging and multimodal communication approaches. Diss. University of Illinois at Urbana-Champaign, 2022.

[8] Ironsi, Chinaza Solomon. "Exploring the potential of generative AI in English language teaching." Facilitating global collaboration and knowledge sharing in higher education with generative AI. IGI Global Scientific Publishing, 2024. 162-185.

[9] Nanduri, Dinesh Kumar. Exploring the Role of Generative Artificial Intelligence in Culturally Relevant Storytelling for Native Language Learning Among Children. MS thesis. University of Maryland, College Park, 2024.