Advances in Social Science and Culture
ISSN 2640-9682 (Print) ISSN 2640-9674 (Online)
Vol. 8, No. 2, 2026
www.scholink.org/ojs/index.php/assc

Original Paper

Effects of Multimodal AI on Emotional Engagement and Cognitive Strategies in Speech Training

Siyu Li[1]

[1] Sichuan Normal University, Chengdu, Sichuan, China

Received: December 29, 2025 Accepted: February 27, 2026 Online Published: March 11, 2026

doi:10.22158/assc.v8n2p58 URL: http://dx.doi.org/10.22158/assc.v8n2p58

Abstract

This study, framed within sociocultural theory, investigates the impact of multimodal AI-assisted speech training on foreign language learners' emotional engagement and cognitive strategies. This research employed a mixed-methods approach to analyze the emotional experiences and cognitive regulation mechanisms for Chinese university students in an AI-assisted speech training environment. Results reveal that learners exhibit moderate-to-high enjoyment and moderate-to-low anxiety in this environment. Foreign language enjoyment and foreign language anxiety are significantly negatively correlated. No significant linear relationship was found between the frequency of AI use and emotional experience. Qualitative analysis indicates that learners actively and critically integrate AI feedback, demonstrating considerable cognitive strategizing and agency. The study identifies limitations of AI in providing emotional support and simulating authentic pressure, especially its difficulty in replacing the emotional interaction and contextual creation functions of human teachers. The results unveil the double-edged sword effect of AI in speech training, underscore the necessity of constructing a human-machine collaborative teaching ecosystem, and provide empirical evidence for the design and integration of intelligent pedagogical tools.

Keywords

Multimodal Artificial Intelligence, Speech Training, Emotional Engagement, Cognitive Strategies, Foreign Language Learning, Human-Machine Collaboration

1. Introduction

1.1 Multimodal AI and Language Learning

In recent years, with the rapid development of artificial intelligence technology, multimodal AI is progressively reshaping the paradigm of language learning and skill training (Chen, Smith, York, & Mayall, 2021). Multimodal AI refers to intelligent systems capable of integrating, processing, and

generating information across multiple modalities, including text, speech, images, and video (Baltrušaitis, Ahuja, & Morency, 2019). In language learning, particularly in the training of productive skills, its application has expanded from basic grammar correction to complex oral communication and speech training. In the domain of speech and presentation training, multimodal AI integrates speech recognition, computer vision, and natural language processing technologies (Wang & Zhao, 2022). These technological applications aim to create a repeatable, low-risk, and highly personalized training environment to compensate for insufficient immediate feedback and limited opportunities for authentic practice in traditional teaching. Existing empirical studies largely confirm the positive role of such technologies in enhancing oral fluency, accuracy, and learner motivation. For instance, systems based on speech recognition and immediate feedback can significantly improve learners' pronunciation accuracy and oral fluency (Li, Chen, & Zhang, 2020). Simultaneously, multimodal platforms integrating visual and speech analysis can enhance learners' expressive confidence and task engagement. However, to our knowledge, existing studies still adhere to a binary "technology intervention – learning outcome" validation model. In-depth analysis of the "black box" issue concerning how technology influences learners' internal emotional states and cognitive processing need more attention.

1.2 Research on Foreign Language Emotions

The role of emotional factors in foreign language learning has become a key area in second language acquisition research. Among these, foreign language anxiety, as a typical negative emotion, has been widely confirmed to have an inhibiting effect on the foreign language learning process and outcomes (Horwitz, Horwitz, & Cope, 1986). Numerous scholars have also focused on the specific manifestations and impacts of foreign language anxiety in high-stakes productive tasks such as public speaking. For example, Woodrow (Woodrow, 2006) found that in impromptu speaking tasks, foreign language anxiety significantly reduces students' oral fluency and syntactic complexity, leading to more avoidance behaviors. Furthermore, Horwitz (2010) pointed out that public speaking anxiety is often accompanied by distinct physiological arousal (e.g., accelerated heartbeat, vocal tension) and cognitive interference (e.g., distraction, disrupted logical flow of expression), thereby affecting overall speech performance and self-efficacy (Bandura, 1997). These studies collectively indicate that foreign language anxiety is prevalent in public speaking contexts and exerts a clear inhibitory effect on language output quality and psychological states. In contrast, research on positive emotions, with foreign language enjoyment at its core, has emerged more recently. Studies in recent years have shown that enjoyment is significantly positively correlated with stronger learning motivation, higher classroom engagement, and better learning outcomes (Dewaele & MacIntyre, 2014).

Among all language tasks, delivering a speech is especially with high emotional arousal. Delivering a speech is a complex task integrating high cognitive load and high emotional arousal (Sweller, 1988). From a cognitive perspective, it involves a series of intensive mental processes, including information organization, linguistic encoding, logical construction, and on-the-spot adaptation. From an emotional

perspective, speakers often find themselves under the social evaluative pressure of public expression, experiencing anxiety, the tension of self-monitoring, and anticipation and worry regarding audience feedback. Therefore, speech training concerns the refinement of linguistic competence and requires learners to flexibly regulate their emotions and cognitive strategies. Anxiety and enjoyment often co-exist and may transform into one another, jointly shaping learners' behaviors and experiences. However, to our knowledge, existing research often treats these two as independent psychological constructs for measurement and reporting, or merely conducts simple comparisons. Their symbiotic, competitive, and transformative relationship within specific, dynamic task execution processes still need more exploration.

1.3 The Present Study

In summary, the existing research exhibits the following limitations: Firstly, there is a lack of systematic investigation into learners' subjective emotional journeys and real-time cognitive strategy use during AI-assisted training. Secondly, the focus on emotions is narrow, overly concentrating on the negative dimension of anxiety, failing to adequately integrate the perspective of positive emotions to form a more comprehensive picture of emotional engagement. Thirdly, there is analytical fragmentation, where emotional variables and cognitive processes are often studied separately, preventing an examination of their interactive effects within a unified technology-mediated activity framework. Finally, there is mechanistic ambiguity regarding the specific psychological pathways through which multimodal AI feedback affects learners' emotional arousal and cognitive regulation, with a lack of deep experiential evidence from the learners' subjective perspective.

Based on the above, this study adopts a mixed-methods approach, taking multimodal AI-assisted speech training as its specific context. It first quantitatively depicts the overall characteristics of learners' emotional engagement (covering both enjoyment and anxiety dimensions). Subsequently, through purposive sampling and in-depth interviews, it seeks to uncover the underlying explanatory mechanisms of how multimodal AI-assisted training influences learners' emotional and cognitive processes. The goal is to build a bridge between technological application and learners' inner psychological world, deepen relevant research, and provide empirical support for AI-assisted speech training.

Specifically, this study proposes two research questions: First, what impact does multimodal AI-assisted speech training have on learners' emotional experiences? Second, what impact does multimodal AI-assisted speech training have on learners' cognitive strategies?

2. Research Methodology

2.1 Participants

70 Chinese university students (15 males, 55 females, aged 19-23) were included in this study. , All participants were native speakers of Mandarin Chinese, with English as their second language. They had an average of 10 years of English learning experience and at least six months of speech training

experience. Concurrently, all participants had undergone speech training courses (approximately two hours per week) and had long-term experience using AI-assisted speech training tools. All participants were right-handed, had no history of mental illness, and possessed normal or corrected-to-normal vision.

2.2 Research Instruments

This study employed a mixed-methods approach combining a questionnaire survey and semi-structured interviews. The questionnaire was used to collect quantitative data, while interviews were conducted with a subset of participants for in-depth understanding, supplementing and enriching the questionnaire results.

2.2.1 Questionnaire Survey

The questionnaire for this study was adapted based on the Foreign Language Classroom Anxiety Scale (FLCAS) by Horwitz et al. and the Foreign Language Enjoyment Scale (FLES) by Dewaele and MacIntyre, with appropriate revisions tailored to the speech training context. The questionnaire utilized a 5-point Likert scale. For positively worded items, responses ranging from "strongly disagree" to "strongly agree" were scored from 1 to 5, respectively. For negatively worded items, responses from "strongly agree" to "strongly disagree" were scored from 5 to 1, respectively. An attention-check item was included, requiring participants to select a specific fixed option to verify attentive responding. Questionnaires failing this check were excluded from the analysis.

2.2.2 Semi-structured Interviews

To gain an in-depth understanding of learners' emotional and cognitive processes during AI-assisted speech training, three participants were randomly selected for semi-structured interviews after completing the questionnaire. These interviews aimed to explore their authentic experiences and psychological changes while using multimodal AI-assisted speech training. The interview protocol was structured around dimensions of emotional experience, cognitive regulation, and perceived differences. It explored immediate emotional reactions to various modalities of AI feedback (e.g., voice evaluation, visual cues) and overall emotional perceptions. It also investigated strategic adjustments made to speech content, delivery style, or psychological state after receiving AI feedback. Furthermore, it compared the differences between AI feedback and feedback from human teachers or audiences, and its impact on self-efficacy and learning motivation. To facilitate clear expression of ideas, interviews were conducted in Chinese. Upon completion, the interview content was transcribed verbatim for further analysis.

2.3 Experiment Procedure

After participants completed a period of multimodal AI-assisted speech training tasks, the questionnaires were distributed. Subsequently, 3 interviewees were invited for one-on-one interviews, each lasting approximately 30-45 minutes. All interviews were audio-recorded and transcribed into textual data.

2.4 Data Analysis Methods

A total of 70 questionnaires were collected. After excluding invalid responses, 65 valid questionnaires remained. SPSS 26 was used to conduct descriptive statistics and correlation analyses on the questionnaire data. The transcribed interview texts were analyzed following the thematic analysis method outlined by Braun and Clarke (2006), involving coding and theme development.

3. Results

3.1 Quantitative Results

The descriptive statistics and correlation analysis results for foreign language enjoyment and anxiety are presented in Table 1. Learners' enjoyment in the multimodal AI-assisted speech training context was generally at a moderately high level (M = 3.82, SD = 0.61), while anxiety was generally at a moderately low level (M = 2.45, SD = 0.73). Correlation analysis indicated a significant negative correlation between enjoyment and anxiety ($r = -0.42$, $p < .01$), suggesting that within this environment, higher levels of enjoyment were often accompanied by lower levels of anxiety.

Table 1. Descriptive Statistics of Foreign Language Enjoyment and Anxiety (N = 65)

Variable	M	SD	Min	Max
Foreign Language Enjoyment	0.80	0.45	-0.13	2.00
Foreign Language Anxiety	0.13	0.52	-1.67	2.00

To further explore emotional differences among learners with varying frequencies of AI use, participants were divided into high, medium, and low usage groups based on their self-reported weekly frequency of using the tool for speech training, employing the tertile split method: Low usage frequency group (≤ 2 times/week), Medium usage frequency group (3-4 times/week), High usage frequency group (≥ 5 times/week). The descriptive statistics for enjoyment and anxiety across these groups are shown in Table 2. Overall, the high usage frequency group exhibited relatively higher mean scores for both enjoyment and anxiety, while the low usage frequency group showed a relatively lower anxiety level.

Table 2. Descriptive Statistics of Foreign Language Enjoyment and Anxiety by AI Usage Frequency Group (M ± SD)

Group	Foreign Language Enjoyment	Foreign Language Anxiety
High Usage Frequency	0.85 ± 0.48	0.30 ± 0.75
Medium Usage Frequency	0.79 ± 0.50	0.02 ± 0.56
Low Usage Frequency	0.79 ± 0.43	0.13 ± 0.47

Correlation analysis results (see Table 3) confirmed a significant negative linear relationship between

foreign language enjoyment and anxiety ($r = -0.42$, $p < .01$), indicating that in the AI-assisted speech training context, the stronger the learners' feelings of enjoyment, the lower their anxiety levels tended to be, demonstrating a degree of emotional concordance.

Table 3. Linear Regression Analysis Results of AI Usage Frequency on Foreign Language Enjoyment and Anxiety

Dependent Variable	Independent Variable	B	*SE*	β	*t*	*p*	R^2
Foreign Language Enjoyment	AI Usage Frequency Group	-0.032	0.049	-0.081	-0.645	.521	.007
Foreign Language Anxiety	AI Usage Frequency Group	-0.057	0.057	-0.125	-0.999	.32	.016

Regression analysis results showed that AI usage frequency did not significantly predict enjoyment levels ($\beta = -0.081$, $p = .521$), with the model explaining a very low amount of variance ($R^2 = .007$). This indicates no significant linear relationship between AI usage frequency and enjoyment experience. Similarly, AI usage frequency did not significantly predict anxiety levels ($\beta = -0.125$, $p = .322$), with the model explaining only 1.6% of the variance ($R^2 = .016$), suggesting no significant linear correlation between AI usage frequency and anxiety experience.

Overall, within the current sample, AI usage frequency had limited explanatory power for individual enjoyment and anxiety emotions, and no significant linear relationship was found between them. This key finding suggests that the core factors influencing emotional experience in AI-assisted speech training may not be simply the "quantity" of use, but rather the "quality" of the interaction between the user and the AI. This includes specific usage strategies, the degree of cognitive engagement, and the ways in which feedback is interpreted and integrated. This provides an important quantitative basis for the subsequent qualitative analysis focusing on interaction processes and subjective experiences.

3.2 Qualitative Analysis Results

The interviews first confirmed the value of AI-assisted speech training in alleviating anxiety and enhancing a sense of security. Specifically, interviewees commonly mentioned feeling "relaxed" and "pressure-free," indicating that AI-assisted speech training effectively mitigated immediate anxiety. However, this sense of security stemmed from the "non-human" nature of the interaction partner, precisely exposing a fundamental deficiency of current AI in simulating authentic socio-emotional support. Furthermore, the interview results indicated that the motivational effect of AI feedback exhibited a clear temporality. Its formulaic pattern of emotional feedback could be encouraging initially, but this motivational effect diminished with increased usage. For example, interviewees generally expressed sentiments like, "At first, I felt encouraged, but later it felt like a fixed routine." This raises deep questions about the sustainability of technology-mediated emotional incentives, suggesting that algorithmically generated "encouragement" alone may struggle to support learners' long-term intrinsic motivation.

Notably, while AI-assisted speech training demonstrated strong practical value at the cognitive level,

with interviewees unanimously appreciating its excellent instrumentality in analyzing objective dimensions such as linguistic fluency and logical structure, what emerged as even more critical within the speech training environment was the high degree of strategizing and agency displayed by the learners themselves (Flavell, 1979; Piaget, 1970; Vygotsky, 1978). For instance, Interviewee A stated, "I first outline my speech myself, then let the AI help me optimize the logic." Similarly, Interviewee B said, "I compare AI and teacher suggestions and choose the more suitable one." This indicates that all participants pragmatically positioned AI as a personal "productivity tool" for learning, whose effectiveness was highly dependent on the user's clear instructions and preliminary adjustments. They were not passive recipients of information but actively engaged in critically screening, cross-tool comparison of AI outputs, and even deepening understanding through deliberate "argumentation" (as described by Interviewee C). This behavioral pattern reveals a key shift: in AI-enhanced learning environments, the core cognitive activity is moving from "information acquisition" to "information evaluation and integration." Learners' metacognitive abilities and critical thinking become crucial for harnessing tools and achieving effective learning. Moreover, due to its inability to simulate real interpersonal pressure, AI-assisted speech training played a minimal role in cultivating the "psychological resilience" required for facing authentic audiences. These findings prompt a reconsideration of AI's role within a complete educational ecosystem. Speech training should arguably apply AI to standardized, repeatable practice segments (e.g., content revision, repetitive drills), while leaving tasks such as emotional support, creative inspiration, and high-fidelity pressure simulation to human mentors or future highly immersive virtual environments. Effective "human-machine collaboration" should be built upon functional complementarity based on respective strengths, rather than simple substitution.

4. Discussion

4.1 The Double-Edged Sword Effect of AI

The quantitative data from this study presented an overall emotional profile characterized by "moderate-to-high enjoyment and moderate-to-low anxiety", while the qualitative data revealed that the core mechanism underlying this profile is technology-mediated "high perceived controllability." The AI environment, by eliminating the threat of social evaluation, directly reduces a key source of foreign language anxiety, enhancing speakers' tolerance for emotional anxiety. This low-anxiety state facilitates the release of cognitive resources, allowing speakers to focus their attention on the task itself. Simultaneously, AI deconstructs complex tasks into micro-steps with immediate feedback, forming a positive cycle of "achieving small goals – obtaining immediate enjoyment" (Locke, & Latham, 2002). However, this controllability is a double-edged sword. While it lowers anxiety, it also strips away the unpredictable interpersonal pressure elements inherent in real-life speeches, potentially leading to a "greenhouse effect" – learners may perform adeptly in the AI environment but remain unprepared for the challenges of authentic contexts. Furthermore, the diminishing motivational effect associated with

formulaic AI feedback also suggests a ceiling effect on the quality and depth of this type of emotional support.

4.2 The Dual Challenge Posed by AI

The findings strongly refute the view of learners as passive recipients of technology, establishing their role as "active harnessers". Learners strategically utilized AI for interdisciplinary research and logical refinement, critically integrating its outputs. This finding resonates with constructivist learning theories and directs attention towards learners' "meta-competencies" within human-machine collaboration. However, the effective exercise of this agency is conditional. It requires learners to possess a certain foundational knowledge base, clear self-awareness (metacognition), and critical thinking skills. For learners lacking these prerequisite skills, the AI's "command-feedback" model may fail to initiate effective learning processes and could even exacerbate their confusion. This poses a dual challenge for educators and technology designers: on one hand, cultivating learners' relevant "meta-competencies"; on the other, exploring how to design AI interaction paradigms that are more guiding and capable of "empowering" lower-proficiency users, for instance, by providing structured question scaffolds or multimodal inspirational materials.

4.3 Towards an AI-Oriented Human-Machine Collaborative Ecosystem

This study clearly reveals a "capability paradox" of current AI in speech training: excellence in information processing and objective analysis, yet weakness in emotional interaction and contextual understanding. This is not a technological failure but an accurate mapping of its capability boundaries. It clearly identifies the irreplaceable core value of human teachers: providing emotional connection, personalized care, improvisational guidance based on rich experience, and the creation of authentic pressure situations (Hargreaves, 1998). Therefore, the future direction should not be the pursuit of replacing teachers with AI, but rather committing to constructing a "human-machine collaborative" intelligent ecosystem. Specific design implications include: developing immersive virtual environments that simulate real-time audience reactions to fill the gap in "pressure training"; optimizing AI's feedback logic so that when user instructions are ambiguous, it can proactively offer optional problem frameworks or case references to lower the barrier to use; and exploring personalized adaptation pathways based on learning analytics, enabling AI to more intelligently adapt to learners' dynamic developmental needs. The ultimate goal is to leverage the respective strengths of AI and human teachers, forming a seamlessly connected, functionally complementary, and comprehensive training system to jointly support learner development.

5. Conclusion

This study employed a mixed-methods approach to investigate learners' emotional and cognitive engagement in multimodal AI-assisted foreign language speech training, drawing the following main conclusions:

Firstly, at the emotional level, by providing a highly controllable, low-risk space, the AI environment fostered an overall emotional experience characterized by "moderate-to-high enjoyment and moderate-to-low anxiety," with a significant negative correlation between the two. AI usage frequency did not show a significant linear relationship with emotional experience, indicating that frequency of use is not a primary predictor of emotional outcomes. However, limitations were observed in the formulaic nature of AI's emotional motivational effects, and the environment's lack of authentic interpersonal pressure simulation may constrain its coaching efficacy.

Secondly, at the cognitive level, learners demonstrated notable high strategic thinking and agency, critically integrating AI into their personal cognitive strategies for content expansion and logical refinement. This process underscores the importance of metacognition and critical thinking as core learning competencies in technology-rich environments.

Finally, at the system efficacy level, AI is confirmed as a powerful "productivity tool" for information processing and objective analysis, but it possesses inherent boundaries in providing deep emotional support, achieving contextualized understanding, and simulating authentic speech pressure. Consequently, its appropriate positioning is as a complementary tool within the human teacher-led pedagogical ecosystem, rather than a substitutive coach.

However, this study also has certain limitations: the relatively small sample size and homogeneous background limit the generalizability of the findings. Future research may expand sample diversity, employ longitudinal tracking or comparative experimental designs, and delve deeper into the differential impacts of various AI interaction modes (e.g., highly immersive virtual audiences, emotional conversational agents) on learning psychological engagement.

References

Baltrušaitis, T., Ahuja, C., & Morency, L.-P. (2019). Multimodal machine learning: A survey and taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *41*(2), 423-443.

Bandura, A. (1997). *Self-efficacy: The exercise of control*. W.H. Freeman and Company.

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology*, *3*(2), 77-101.

Chen, Y., Smith, T. J., York, C. S., & Mayall, H. J. (2021). The effects of a multimodal AI feedback system on speaking performance and self-efficacy in an EFL context. *Computer Assisted Language Learning*, *34*(5-6), 567-589.

Dewaele, J.-M., & MacIntyre, P. D. (2014). The two faces of Janus? Anxiety and enjoyment in the foreign language classroom. *Studies in Second Language Learning and Teaching*, *4*(2), 237-274.

Flavell, J. H. (1979). Metacognition and cognitive monitoring: A new area of cognitive-developmental inquiry. *American Psychologist*, *34*(10), 906-911.

Horwitz, E. K. (2010). Foreign and second language anxiety. *Language Teaching*, *43*(2), 154-167.

Horwitz, E. K., Horwitz, M. B., & Cope, J. A. (1986). Foreign language classroom anxiety. *The*

Modern Language Journal, *70*(2), 125-132.

Li, C., Chen, T., & Zhang, J. (2020). The impact of automated speech recognition feedback on pronunciation accuracy and fluency in second language speaking. *Language Learning & Technology*, *24*(2), 54-73.

Locke, E. A., & Latham, G. P. (2002). Building a practically useful theory of goal setting and task motivation: A 35-year odyssey. *American Psychologist*, *57*(9), 705-717.

Piaget, J. (1970). *Science of education and the psychology of the child* (Trans. D. Coltman). Viking.

Sweller, J. (1988). Cognitive load during problem solving: Effects on learning. *Cognitive Science*, *12*(2), 257-285.

Vygotsky, L. S. (1978). *Mind in society: The development of higher psychological processes* (M. Cole, V. John-Steiner, S. Scribner, & E. Souberman, Eds.). Harvard University Press.

Hargreaves, A. (1998). The emotional practice of teaching. *Teaching and Teacher Education*, *14*(8), 835-854.

Wang, Y., & Zhao, L. (2022). Enhancing public speaking confidence and engagement through a multimodal AI training platform: An empirical study in a Chinese EFL context. *System*, *108*, 102841.

Woodrow, L. (2006). Anxiety and speaking English as a second language. *RELC Journal*, *37*(3), 308-328.