Proceedings of the 4th International Conference on Art, Design and Social Sciences
DOI: 10.54254/2753-7064/2025.25788

Multimodal Affective Sensing Driven Adaptive Intervention Model for Second Language Learning Stress

Yixin Hua

University of Melbourne, Melbourne, Australia
yixinhua2002@outlook.com

Abstract. In high-pressure learning contexts, second language (L2) learners often struggle with the interplay between anxiety and cognitive overload, significantly impairing their language acquisition efficiency. However, most existing intelligent education systems fail to dynamically detect and regulate learners' psychological states throughout the learning process. To address this gap, this study proposes an adaptive intervention model that integrates multimodal emotion recognition with reinforcement learning-based strategy optimization. By capturing learners' facial expressions, vocal prosody, and physiological signals at key task stages, the system enables real-time emotional state recognition and generates personalized intervention strategies based on emotional feedback. Sixty university-level English learners were recruited for an 8-week randomized controlled trial. While the control group followed a conventional instructional approach, the experimental group was supported by the emotion-sensing intervention system. The study employed standardized language tests and system performance metrics to assess the effectiveness of the intervention, supplemented by learning logs and interviews to collect subjective user feedback. Results show that the experimental group outperformed the control group in terms of emotion recognition accuracy, intervention response latency, and language performance improvement. Learners also reported high acceptance and positive evaluations of the system. This research validates the feasibility of multimodal affective sensing in mitigating learning stress and provides both technical and empirical foundations for emotion-adaptive intelligent education systems, thereby expanding the application boundary of affective computing in intelligent instruction.

Keywords: Multimodal sensing, Emotion recognition, Learning stress, Personalized intervention, Second language acquisition

1. Introduction

In a globalized context, second language acquisition is not only a cognitive endeavor but also an emotionally charged process. Under high-pressure learning conditions, learners frequently experience cognitive overload, emotional fluctuation, and motivational decline, all of which undermine the efficiency of language input and output [1]. Even though cognitive education systems have made significant work about content adaptation and the suggestion of learning paths, none have ever considered the dynamic emotional states of learners. Along with the advancement of affective computing and multimodal sensing technologies, the possibilities emerge of extracting emotional features when learners are learning something and performing personalized interventions accordingly [2]. According to the above, the current work proposes a multimodal affective sensing-adaptive intervention model that fuses the use of facial expressions, speech prosody, and body-attached physiological signals for the real-time extraction of L2 learning stress as well as offering personalized help. Trying to enhance learners' emotional regulation and their acquisition performances as well, the model enables a paradigm shift for the intelligent education systems from the knowledge-centrism of adaptation to sensitivity to emotions.

2. Literature review

2.1. Applications of multimodal affective sensing in education

The core of multimodal affective sensing is the integration of data streams of various sensors to establish a thorough learner psychological state profile. Most existing studies use facial expression analysis, speech feature extraction, and physiology signal analysis as input modalities, integrating them with the help of deep neural networks to improve the accuracy of emotional categorization [3]. However, the studies are mainly static emotion detection and lack temporal modeling, thus incapable of reflecting nonlinear dynamics of emotional variation for the purpose of education. Moreover, while the accuracy of identifying emotions has been enhanced, the practical linkage of affective sensing with the process of education intervention yet lacks thorough investigation [4]. Building on the strengths of multimodal fusion, this study introduces temporal modeling and feedback coupling mechanisms to transition emotion recognition systems from passive detection to active response, aiming to embed them effectively into instructional practices.

2.2. Cognitive and emotional effects of stress on language learning

Interdisciplinary research between cognitive psychology and affective neuroscience reveals that learning stress interferes with language processing in multiple ways. These include restricted attentional resource allocation, reduced working memory capacity, and impaired emotional regulation, which collectively hinder phonological discrimination, vocabulary retention, and syntactic processing [5]. Further studies indicate that learners under prolonged stress exhibit deficits in language transfer, strategy use, and intercultural communication. Traditional language instruction often overlooks the impact of stress on performance, relying solely on teachers' subjective judgment or test results, which leads to delayed and ineffective interventions [6]. Therefore, the ability to accurately detect and respond to learning stress is critical for improving L2 instruction. Through emotion modeling and data-driven analysis, this study aims to reconstruct the dynamic linkage between emotion, cognition, and language performance.

2.3. Strategies for emotion-adaptive intervention

The essence of an adaptive intervention system lies not in the abundance of pre-defined content but in its capacity to dynamically adjust intervention paths based on learners' real-time states, thereby enabling precise coordination among the human, machine, and learning context. Existing systems largely rely on rule-based mechanisms or fixed templates, which are ill-suited for handling emotional variability and learner heterogeneity in complex educational settings [7]. With recent advances in reinforcement learning, generative models, and interactive systems, emotion-driven intelligent intervention has become a key research focus. Such systems emphasize that the generation of feedback strategies must align with the learner's current emotional state, frequent behavior patterns, and long-term goals to produce continuous and context-sensitive responses [8]. By leveraging multimodal inputs, temporal signal analysis, and personalized modeling, this study constructs a strategy decision model capable of real-time feedback optimization, advancing the system from predictive reaction to generative support.

3. Methodology

3.1. Participants and experimental design

This study recruited 60 university-level English learners as participants. Selection criteria required that learners possess a basic level of English proficiency and be under academic or exam-related learning pressure. Participants were randomly assigned to either the experimental or control group, with 30 learners in each. The experiment lasted eight consecutive weeks and covered a complete cycle of English skill training and assessment. All the students followed the same course activities and materials; the experimental group, though, had the aid of a multimodal adaptation-based emotion-driven intervention system, while the control group followed regular classroom instructions without real-time emotional perception and feedback, which served as the control group.

A mixed-methods design was adopted. For the quantitative component, standardized language tests were used to measure vocabulary acquisition, grammar accuracy, and listening comprehension before and after the intervention. System-level indicators such as emotion classification accuracy and intervention response latency were also recorded. On the qualitative side, semi-structured interviews and learner diaries were used to collect participants' subjective feedback on the intervention experience, including perceived emotional changes and satisfaction with the system's responsiveness.

3.2. Data collection and processing

Data were collected at three seminal points: the pre-task preparation phase before the onset of the task, at peak cognitive-load periods while the task was being completed, and following the task self-report phase. A camera pointed at the operator provided the system with facial key points to acquire micro-expressions, a speaker to capture speech frequency and shifts in prosody to measure anxiety, and wearable units to capture the following physiological signals: heart rate variability and electrodermal activity [9]. Together, these inputs created a tri-channel system with visual, auditory, and bodily data.

To ensure temporal alignment and data consistency, the system adopted a unified sampling rate (30Hz for video, 16kHz for audio, and 1Hz for heart rate) and a fixed data window length of 10 seconds. Raw data were first processed through a preprocessing module involving noise reduction, normalization, dimensional standardization, and outlier removal [10]. The cleaned data were then input into a Transformer-based deep multimodal fusion network for feature extraction and emotional state prediction. The system classified learners' emotional states into seven categories—calm, focused, anxious, fatigued, stressed, frustrated, and discouraged—which served as the basis for adaptive intervention selection and feedback path planning.

3.3. Intervention strategy modeling

After recognizing learners' emotional states via multimodal sensing, the system entered the strategy decision and implementation phase. To address personalized intervention needs, the study constructed a strategy library comprising various forms of interventions, including text prompts, audio guidance, visual cues, and cognitive load regulation [11]. The system dynamically selected intervention paths using a strategy-response pairing mechanism. A reinforcement learning (RL) algorithm was used to build the intelligent decision module, with core optimization performed using the Q-learning method. The update rule is as follows:

$$Q\left(s_t, a_t\right) \leftarrow Q\left(s_t, a_t\right) + \alpha\left[r_{t+1} + \gamma \max_a Q\left(s_{t+1}, a\right) - Q\left(s_t, a_t\right)\right] \tag{1}$$

Here, s_t represents the current state (e.g., "anxious + timed test"), a_t is the current intervention action (e.g., "play guided meditation"), and r_{t+1} is the immediate reward (e.g., reduced heart rate or extended focus time). α is the learning rate, and γ is the future reward discount factor.

To further enhance strategy robustness and generalization, a policy gradient method was employed to optimize the policy function π(a ∣ s), with the update defined as:

$$\nabla J\left(\theta\right) = E_{\pi_\theta}\left[\nabla_\theta \log \pi_\theta\left(a|s\right) \cdot R\right] \tag{2}$$

Where θ denotes the policy parameters and R the accumulated reward signal. This mechanism allows the system to continually optimize feedback strategies based on long-term emotion-behavior trajectories, facilitating a shift from reactive intervention to predictive regulation.

4. Results

4.1. Performance of emotion detection and feedback system

The multimodal emotion recognition system achieved an average classification accuracy of 87.3% across the seven primary emotional states, with "anxious" having the highest accuracy (92.1%) and "fatigued" the lowest (81.4%). Response latency across all states was below 2 seconds, with an average of 1.47 seconds, meeting the requirements for real-time intervention. Further analysis showed that facial expression features contributed the most (43.2%) to recognition accuracy, followed by vocal prosody (35.8%) and physiological signals (21.0%), as shown in Table 1.

Table 1. Analysis of data

Emotion State	Accuracy (%)	Response Time (s)	Facial Weight (%)	Voice Weight (%)	Physiological Weight (%)
Calm	88.7	1.42	41.3	37.2	21.5
Focused	89.5	1.35	44.8	33.9	21.3
Anxious	92.1	1.28	47.6	36.4	16
Fatigued	81.4	1.75	38.9	35.1	26
Stressed	86.8	1.53	42.7	37.8	19.5
Frustrated	84.2	1.68	40.1	36.5	23.4
Discouraged	85.9	1.52	44.1	34.2	21.7
Average	87.3	1.47	43.2	35.8	21

In high cognitive-load tasks such as timed listening comprehension, emotion recognition accuracy increased by 4.7%, primarily due to the heightened emotional expressiveness under stress. The reinforcement learning algorithm demonstrated strong convergence, with the Q-value function stabilizing after approximately 2,000 iterations. The policy gradient method further improved cumulative rewards by 12.8%.

4.2. Impact on learning outcomes

In vocabulary tests, the experimental group's post-test scores (mean: 78.4) improved by 16.3 points from the pre-test (mean: 62.1), a 26.2% increase. In contrast, the control group improved by only 9.7 points (17.1%). For grammar accuracy, the experimental group improved from 64.8% to 82.3%, an increase of 17.5 percentage points, while the control group increased from 65.1% to 76.9%, an 11.8-point gain. Listening comprehension showed the most significant improvement, with the experimental group improving from 59.3 to 81.7 points (37.8%), compared to the control group's improvement from 58.9 to 71.2 points (20.9%), as illustrated in Figure 1.

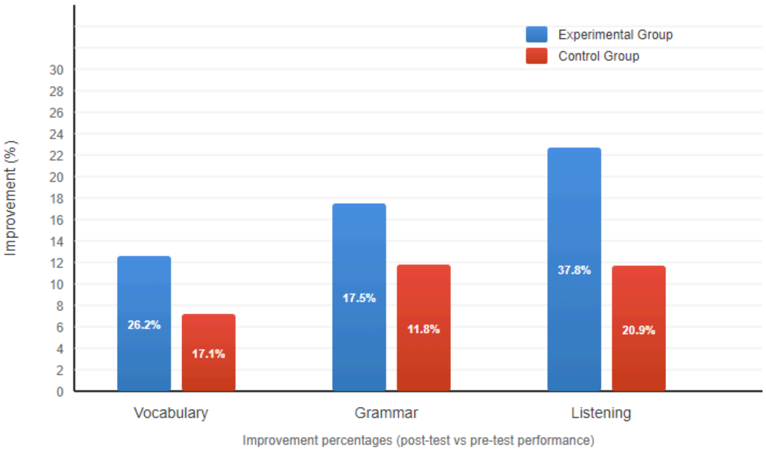

Figure 1. Learning outcomes comparison: experimental vs. control groups

Statistical analysis confirmed that differences between the groups in all test areas were significant ($p < 0.05$). The experimental group performed particularly well under high-pressure tasks. In timed vocabulary tests, their average response time was 14.2% shorter, and error rates were 19.7% lower than the control group. In grammar cloze tasks, the experimental group outperformed the control group by 23.1% in complex structure identification accuracy. Listening comprehension sub-analysis revealed that the experimental group scored significantly higher across detail recognition, inference, and main idea extraction, with inference showing the greatest improvement—an average of 4.8 points higher than the control group.

5. Discussion

The results of this study demonstrate that the multimodal affective sensing–driven adaptive intervention model significantly enhances the identification of learning stress, optimizes feedback responsiveness, and improves language acquisition outcomes. By integrating emotional state monitoring into the instructional feedback loop, the system achieved precise detection of emotional fluctuations under pressure and enabled personalized strategy generation. This mechanism offers a new paradigm for AI-powered educational systems, wherein feedback is informed not only by performance metrics but also by real-time psychological dynamics. The study involved a small, context-specific sample, which may limit generalizability. Emotion labeling was affected by signal noise and subjectivity. Future work should expand sample diversity, refine labeling methods, and enhance cross-modal modeling to validate system robustness and applicability.

6. Conclusion

This study developed and validated an adaptive intervention model based on multimodal emotion recognition and reinforcement learning, enabling real-time detection of emotional states in high-pressure language learning contexts and providing personalized feedback. Experimental results confirmed that the model significantly improved both emotional regulation and learning outcomes. The findings contribute to the transition of intelligent education systems from purely cognitive adaptation to emotion-cognition co-regulation. Future work should focus on extending the model to broader learner populations, refining long-term intervention mechanisms, and enhancing the system's practical deployment in real-world educational environments.

References

[1] Immadisetty, Praneeta, et al. "Multimodality in online education: a comparative study." Multimedia Tools and Applications (2025): 1-34.

[2] Guo, Xiaoshuang. "Multimodality in language education: implications of a multimodal affective perspective in foreign language teaching." Frontiers in Psychology 14 (2023): 1283625.

[3] Alqarni, Nada A. "Predictors of foreign language proficiency: Emotion regulation, foreign language enjoyment, or academic stress?." System 126 (2024): 103462.

[4] Tanaka, Hiroki, et al. "4th Workshop on Social Affective Multimodal Interaction for Health (SAMIH)." Proceedings of the 25th International Conference on Multimodal Interaction. 2023.

[5] Govea, Jaime, et al. "Implementation of deep reinforcement learning models for emotion detection and personalization of learning in hybrid educational environments." Frontiers in Artificial Intelligence 7 (2024): 1458230.

[6] Yan, Lixiang, et al. "Scalability, sustainability, and ethicality of multimodal learning analytics." LAK22: 12th international learning analytics and knowledge conference. 2022.

[7] Pathirana, Amod, et al. "A Reinforcement Learning-Based Approach for Promoting Mental Health Using Multimodal Emotion Recognition." Journal of Future Artificial Intelligence and Technologies 1.2 (2024): 124-142.

[8] Rahman, Fathima Abdul, and Guang Lu. "A Contextualized Real-Time Multimodal Emotion Recognition for Conversational Agents using Graph Convolutional Networks in Reinforcement Learning." arXiv preprint arXiv: 2310.18363 (2023).

[9] Devulapally, Naresh Kumar, et al. "AM^ 2-EmoJE: Adaptive Missing-Modality Emotion Recognition in Conversation via Joint Embedding Learning." arXiv preprint arXiv: 2402.10921 (2024).

[10] Wang, Zhuozheng, and Yihan Wang. "Emotion recognition based on multimodal physiological electrical signals." Frontiers in Neuroscience 19 (2025): 1512799.

[11] Zhao, Jiaxing, Xihan Wei, and Liefeng Bo. "R1-omni: Explainable omni-multimodal emotion recognition with reinforcement learning." arXiv preprint arXiv: 2503.05379 (2025).