Journal of International Education and Development
2025, VOL. 9, NO. 8, 79-82
DOI: 10.47297/wspiedWSP2516-250018.20250908

Research on the Optimization of a Vocational English Listening and Speaking Ability Training Model Based on AI Technology

Fukai Zhang

School of Humanities and Education, Yancheng Kindergarten Teachers College, Yancheng, Jiangsu, 224000, China

ABSTRACT

Against the backdrop of accelerated digital transformation in vocational education, industrial upgrading has placed new demands on the foreign language proficiency of composite talent. To address the structural challenges in English listening and speaking instruction at higher vocational colleges, this study explores a systematic integration scheme for artificial intelligence technology. Focusing on core issues such as insufficient authenticity of the teaching context, mismatch between technological tools and teaching objectives, and the limited dimensions of assessment mechanisms, this study proposes an optimized framework comprising intelligent reconstruction of vocational contexts, multi-source data-driven training, and dynamic tracking of three-dimensional competencies. The research indicates that this model effectively enhances the practicality and industry relevance of language training by deeply aligning technological characteristics with teaching principles, thereby providing a feasible, practical pathway for the digital transformation of foreign language proficiency development in vocational education.

KEYWORDS

AI technology; Vocational English; Training model

1 Introduction

The global restructuring of industrial chains has placed unprecedented demands on the cross-cultural language skills of technical and skilled personnel. The ability to respond instantly in foreign language service scenarios has become a core competency for modern vocational education talent. Currently, higher vocational English education faces severe challenges, such as traditional listening and speaking training being limited to static scenario simulations, which struggle to align with the rapidly evolving industry context; the lag in teaching feedback conflicts fundamentally with the timeliness of language acquisition, causing the phenomenon of speech errors to become increasingly entrenched. The maturation of artificial intelligence technology has created new possibilities for addressing these challenges. Speech recognition technology surpasses the feedback limits of human teachers in terms of response speed, while natural language processing's contextual restoration capabilities can create high-fidelity professional scenarios. These technological features align closely with the core principle of vocational education—"learning by doing"—providing an engineered implementation path for contextualized teaching. However, the application of these technologies still faces significant obstacles. Existing intelligent tools generally fail to meet the needs of real-world work scenarios, and there is a substantial mismatch between platform functionality and instructional objectives. Evaluation systems overly emphasize superficial language metrics while neglecting critical dimensions such as strategic application. These contradictions highlight the urgency of systemic reform. Against the backdrop of evolving industrial demands and mature technological conditions, there is an urgent need to achieve a paradigm-level integration of artificial intelligence and language acquisition principles through educational technology innovation [1].

2 The foundational suitability of AI technology in vocational English listening and speaking instruction

2.1 The feasibility mechanism of technology-enabled empowerment

Speech recognition technology breaks through the time barriers of traditional feedback. When learners deviate from standard pronunciation, the system instantly locates the deviation point through acoustic feature comparison and triggers an intervention signal before the error pattern forms muscle memory. This millisecond-level response mechanism compresses the time gap between the occurrence of pronunciation errors and the reception of correction signals, embedding the correction process within the neural encoding window of speech production and blocking the path of error solidification at its source. The adaptive advantages of natural language processing are reflected in its dynamic context restoration capabilities. The system analyzes intent cues and emotional tendencies in conversations to generate real-time interaction strategies that comply with occupational scenario rules. Its strength lies in simulating responses to unexpected situations in real-world communication, such as topic shifts or turn-taking disputes, thereby endowing virtual training with the cognitive pressure characteristics of real-world work scenarios. The technical feasibility lies in converting complex communication scenarios into computable interaction logic units. The synergy of the two technologies establishes a closed-loop error correction mechanism: the speech layer instantly diagnoses pronunciation deviations. In contrast, the semantic layer dynamically adjusts contextual pressure, forming a dual intervention system. This layered processing architecture enables precise technical intervention aligned with the multidimensional needs of language acquisition, laying the engineering foundation for educational model transformation [2].

2.2 Characteristics of higher vocational education

The essential characteristics of higher vocational English education lie in the dual overlapping effects of the dynamic nature of the industry context and the cognitive heterogeneity of the teaching subjects, which are reflected in the structural mismatch between the ever-changing demands of the workplace and the fixed teaching resources, as shown in Table 1 below.

Table 1 Characteristics of higher vocational education

Characteristic dimensions	Dynamic industry context	Cognitive differences
Actual performance	Frequent verbal incidents	Subject background determines language cognition patterns
Capability Development Goals	Pragmatic Adaptability Under Pressure	Expression Strategies Suitable for Different Ways of Thinking
Limitations of traditional teaching	Fixed dialogue templates lack realistic pressure situations	Uniform content ignores differences in cognitive structures
Focus of intelligent technology	Analyze real-time conversation characteristics	Build personalized cognitive training sequences

In real-world work scenarios, language interaction is highly uncertain. For example, sudden emotional interruptions from clients or the need to convey emergency instructions during equipment malfunctions require learners to maintain linguistic accuracy under cognitive overload. This ability essentially involves the dynamic allocation of working memory resources. Existing textbooks' predefined fixed dialogue patterns struggle to simulate communication strategies under real-world pressure, such as turn-taking in emergencies or face-saving mechanisms in conflicts. Cognitive differences among learners further complicate instruction, such as engineering students relying on procedural language structures, while service-oriented students need to master high-context communication skills. These differences stem from long-established cognitive patterns, requiring the instructional system to adjust the mental depth of training content dynamically. The value of intelligent technology lies in capturing learners' micro-behavioral characteristics (such as sentence pause locations, self-correction frequency, and turn-taking duration), thereby establishing training pathways tailored to individual cognitive characteristics.

3 Analysis of the challenges in the development of vocational English listening and speaking skills

3.1 Misalignment in the process of technological application

The current teaching environment suffers from systemic biases, primarily manifested in the following areas: the acoustic design of training spaces does not reflect the real-world workplace sound environment, making it difficult for students to adapt to the auditory challenges of the workplace; the teaching system separates language training from physical movement coordination, neglecting the need for multitaskingmultitasking integration of verbal instructions, equipment operation, and visual monitoring in real-world professional scenarios; standardized environments filter out workplace-specific stress factors, making pre-set linear dialogue processes unable to handle sudden communication conflicts. The essence of these deficiencies lies in the fact that the teaching environment has detached language skills from their practical foundations, reducing listening and speaking skills to abstract training detached from tool operation, environmental interaction, and social roles. As a result, while students may master classroom knowledge, they struggle to translate it into problem-solving abilities in real-world work scenarios.

3.2 Discontinuity in the application of technology

There is a misalignment between the functional positioning of existing technical tools and the core competencies required for professional skills. Mainstream speech recognition systems simplify language training into atomized pronunciation correction, and their isolated sentence analysis mode disrupts the three core elements of workplace communication: the continuity requirements of turn-taking, the adaptive strategies for topic shifts, and the repair mechanisms for communication interruptions. When virtual agents engage in multi-round professional dialogues, their pre-programmed response patterns fail to simulate the strategic adjustments necessary in real-world scenarios, thereby hindering learners' ability to gain decision-making experience in handling unexpected situations. The one-sided nature of technological intervention causes the system to overly focus on superficial speech features while neglecting three implicit abilities that determine professional effectiveness: emotional load recognition (judging the customer's emotional state through paralinguistic features), cultural schema adaptation (such as decoding indirect instructions in high-context cultures), and crisis resolution strategies (using vague limiting language to reduce commitment risks). This dimensional deficiency limits technological applications to the level of sensory-motor training, failing to activate the pragmatic decision-making network dominated by the dorsolateral prefrontal cortex. As a result, learners may accurately pronounce words but frequently commit pragmatic errors when engaging in cross-cultural technological negotiations, leading to a decline in the return on investment between technological investment and actual outcomes [3].

3.3 Limitations of the assessment mechanism

The imbalance between construct validity and criterion validity in the assessment indicator system leads to systematic distortions in the diagnosis of language proficiency. Current evaluations place excessive emphasis on quantifiable elements such as speech fluency and vocabulary accuracy. In contrast, implicit dimensions that determine workplace communication effectiveness—such as emotional management skills, conflict resolution techniques, and cultural adaptation strategies—are overlooked due to the complexity of measurement. This indicator bias leads to severe teaching orientation deviations, such as learners mechanically optimizing pronunciation clarity while neglecting the cultivation of professional competencies like emotional factors. The consequence is that trainees exhibit capability dissociation in high-pressure scenarios: their language forms are correct, yet their communication effectiveness is low. Additionally, the absence of a dynamic assessment mechanism results in stage-based tests functioning like low-frequency sampling of a continuous growth process, failing to capture critical transition points in capability development. This discrete assessment leads to several issues: personalized interventions lacking a data reference framework, an inability to identify ability bottlenecks promptly, and teaching adjustments that degenerate into trial-and-error based on experience. Especially for dimensions like speech strategy abilities that develop progressively, their evolutionary trajectories exhibit non-linear characteristics, and static snapshot-style assessments inevitably cause progress signals to be obscured. This assessment model fundamentally conflicts with the ecological development patterns of language abilities, reducing the teaching process to a black-box operation.

4 AI-based pattern optimization path

4.1 Building an intelligent situational feedback system

To address the dual shortcomings of static teaching environments and feedback delays, an intelligent situational feedback system can be built using three technical approaches: scenario simulation, multitask coordination, and stress loading. This system can resolve the structural disconnect between teaching environments and real workplaces: to address acoustic environment distortion issues, typical workplace acoustic characteristics can be dynamically restored, and environmental adaptability can be enhanced through adjustable noise parameters. To address the issue of sensory coordination fragmentation, a real-time interconnection mechanism can be established between voice command reception, device operation response, and visual monitoring verification to reconstruct the multitask processing loop in professional scenarios; to address the issue of missing stress factors, composite sudden variables can be dynamically injected to activate emergency decision-making and conflict resolution capabilities. By constructing a three-dimensional practical framework encompassing physical environment interaction, tool operation execution, and social role assumption, language training is re-anchored to the foundational skills of the workplace, thereby fully integrating skill conversion pathways [4].

4.2 Building a dual-track collaborative training system

To address the disconnect between technical application and professional skill development, a three-dimensional, dual-track collaborative training model can be established. The foundational skill training module uses intelligent acoustic analysis technology to real-time capture pronunciation deviations, providing visual feedback on oral movements and immediate corrections to enhance the accuracy and fluency of language expression, thereby addressing the issues of mechanical repetition and delayed feedback in traditional training methods. The practical decision-making training module dynamically analyzes the logical relationship network within conversations, tracks the evolution of topics and connection points across multiple rounds of interaction, and automatically locks onto key information nodes when encountering sudden topic shifts, generating contextual connection strategies to eliminate the risk of contextual disconnects completely. The cultural adaptation enhancement module continuously scans for emotional fluctuations in conversations. When identifying negative emotions such as confusion or resistance, it generates communication strategy proposals tailored to specific cultural contexts. For example, in high-context cultural scenarios, it automatically provides customized suggestions for indirect expressions and polite wording. This design breaks through the limitations of single-aspect pronunciation training, establishing a comprehensive capability development system encompassing language accuracy, logical adaptability, and cultural negotiation. It deeply integrates technical intervention into the full spectrum of professional communication capabilities, achieving seamless transition from classroom training to real-world workplace application.

4.3 Innovating a dynamic assessment system

To address the limitations of assessment mechanisms, an innovative dynamic assessment system can be established. By capturing neural activity characteristics in real time during language decision-making processes, a continuous tracking mechanism for the evolution of strategic abilities can be established, enabling the system to identify critical states in the development of key skills. When learners handle complex communication scenarios, the assessment module monitors changes in activation patterns in the brain's higher cognitive regions, with a particular focus on the neural representation of strategic flexibility. When conflict resolution abilities transition from mechanical responses to flexible processing, the prefrontal decision-making network exhibits characteristic activity reorganization. This dynamic allocation of neural resources marks the biological foundation of a qualitative change in ability. The dynamic nature of assessment lies in precise intervention during the learning sensitive period. When the system detects that the application of specific strategies has become rigid (such as repeatedly using stereotypical expressions), it immediately initiates an intensive training protocol. Targeted scenarios can be designed that require ambiguous expressions in negotiation situations, forcing learners to break out of their existing thought patterns. The scientific principle behind this intervention lies in leveraging the window of neural plasticity to promote the functional reconnection of decision-making pathways through the creation of cognitive conflict scenarios [5].

5 Conclusion

This study systematically reconstructs the teaching model for listening and speaking in vocational English education using AI technology. It builds an intelligent contextual feedback system based on the dynamic characteristics of occupational scenarios, combines neural sensitive window intervention to prevent the solidification of pronunciation errors, establishes a dual-track data-driven mechanism to elevate the application of technology to the dimension of decision-making ability cultivation, and innovates a dynamic assessment system. Through the identification of critical states in ability development and targeted intervention during learning sensitive periods, it can achieve precise synchronization between teaching adjustments and neural adaptation patterns. This model fundamentally bridges the gap between vocational skill requirements and the functional limitations of technological tools, establishing a new educational ecosystem centered on industry practice.

Funding

2025-2026 Jiangsu Vocational Education Research Project: "Innovative Practice Research on Higher Vocational English Teaching Based on Artificial Intelligence Technology" (NO: XHYBLX2025273); 2024 Yancheng Kindergarten Teachers College Research Project: "AI Technology Empowering Higher Vocational English Learning Research" (NO: ycyz-2024-17)

About the Author

Fukai Zhang, Master's Degree, Lecturer, Research direction : English Teaching, Language Acquisition.

References

[1] Guo Yirong. Exploring a Hybrid Teaching Model for Vocational English Education Empowered by Digital Intelligence [J]. Academic Weekly, 2025, (22): 119-122.

[2] Guo Haiyan. A Study on Integrated Teaching Strategies for Vocational English Education Oriented toward the Development of Cross-Cultural Communication Skills: A Mixed Study Based on Needs Analysis and Classroom Observation [J]. Modern Vocational Education, 2025, (20): 153-156.

[3] Liu Shanshan, Guo Yaojun. A Discussion on the Innovation of English Teaching in Higher Vocational Education under the Context of Digital Transformation [J]. Path to Success, 2025, (19): 17-20.

[4] Zhu Jingye. An Analysis of the Reform Path of English Teaching in Higher Vocational Education Based on the Foreign Language Golden Course Team Competition [J]. Academic Weekly, 2025, (21): 10-12.

[5] Zhang Zhuo, Cong Huihui, Zhang Dingwen. Research on the Construction and Practice of an Artificial Intelligence-Assisted Vocational English Teaching Model [J]. Modern Vocational Education, 2025, (19): 169-172.