Journal of Network and Innovative Computing
ISSN 2160-2174 Volume 14 (2026) pp. 1-11

Article

Analyzing the Promoting Effect of Virtual Reality English Teaching on Students' Oral Expression Skills Using Multimodal Data Fusion Technology

Qijun Zhao *

Zhaotong University, Zhaotong, Yunnan, 657000, China; zhaoqijuntongguo@163.com

Abstract: The purpose of this study is to explore the application of virtual reality technology in English speaking teaching and its facilitating effect on students' oral expression ability. By constructing a multimodal data fusion analysis framework, it systematically evaluates the effects of virtual context on oral fluency, phonological accuracy and pragmatic competence. The study adopted a quasi-experimental design, selecting eighth grade students from two junior high schools in a city, with the experimental group conducting speaking training in a virtual reality environment and the control group using traditional multimedia teaching. The results show that students in the experimental group significantly outperformed the control group in terms of oral fluency, phonological accuracy, and pragmatic competence, especially the frequency of pragmatic errors in cross-cultural scenarios was significantly reduced. The multimodal data fusion technique captured the synergistic patterns of gestures and speech pauses, revealing the deeper mechanisms of virtual situations on language acquisition. A nonlinear relationship between technology acceptance and learning effectiveness was also found, confirming the moderating role of embodied cognition in virtual language learning. In the future, it is necessary to further optimize the equipment and instructional design, and conduct a large-scale tracking study to verify the long-term effects of technological interventions.

Keywords: virtual reality; spoken English teaching; multimodal data fusion; learning motivation; embodied cognition

1. Introduction

1.1. Background of the Study

At present, one of the common problems in the process of university English classroom teaching is the lack of a real English context atmosphere, the knowledge accumulated by students during the class period can only be practiced in the next class, and the class time of university English courses is also very limited [1-3]. In addition, students can only rely on imitating the teacher or the corresponding audio to learn spoken English, completely unable to enjoy the fun of oral communication, which also led to the reduction of students' interest in oral learning, and over time, many students are also missing the courage and ability to communicate in spoken English [4-6]. And with the application of immersive technology in education, virtual reality (VR) technology has become an effective way to solve the above problems.

VR technology is an emerging technology that is gradually gaining popularity, which provides users with a simulated environment by simulating the way of human perception and activity [7-8]. In English teaching, VR technology has begun to be widely used, which creates a more realistic learning environment for students and provides a more intuitive and vivid English learning experience [9-11]. VR technology plays an important role in promoting students' English oral expression [12]. Traditional oral English teaching mainly relies on the teacher's verbal explanation, and students mainly exercise their oral skills by imitating the teacher [13-14]. However, there are many problems with this teaching method,

such as students often cannot get timely feedback, do not have enough opportunities to practice and rehearse, and therefore have difficulty in overcoming problems in oral learning [15-17]. By leveraging VR technology, a realistic English context can be created. For instance, scenes like "American shopping malls" or "streets of London, UK" can be simulated. Students can communicate with other English users in the virtual environment and enhance their oral expression skills [18-21]. In addition, the application of multimodal data fusion technology will further promote this learning effect. In English teaching with VR technology, multimodal data fusion technology is used to build a learning environment with immersion by combining visual, auditory, and other multidimensional data, and to further promote students' oral expression ability by analyzing their expression ability and proposing effective interventions for them [22-25]. In English teaching, the application of VR technology brings an immersive teaching experience for teachers and students, which provides a favorable learning environment for improving students' learning interest, participation, and learning effect in learning spoken English. Literature [26] introduces VR technology and its application in English speaking training, and conducts experiments on students in the School of Foreign Languages of a university of science and technology, and the results show that VR technology significantly improves students' oral expression ability. Literature [27] reviewed the research on the application of VR technology in English oral communication and pointed out that the advantages of VR technology, such as immersion, interaction and feedback, help to improve oral expression ability. Literature [28] examined the impact of the application of VR technology in English teaching on students' oral proficiency, and verified the effectiveness of the technology in improving students' English oral proficiency, and put forward some teaching suggestions to in helping students improve their oral proficiency. Literature [29] reported a study on the effect of VR technology on English language learners' willingness to communicate and speaking proficiency, and pointed out through a comparative experiment that VR technology improved students' general knowledge, motivation, cultural awareness and self-confidence, and contributed to the improvement of oral expression. Literature [30] examined rural secondary school students' perceptions of the use of VR technology in English language teaching, especially in oral skills courses, and the findings indicated that the interactive experience brought by VR technology increased students' motivation and interest in learning. Literature [31] investigated the role of the learning method of virtual chat (VRChat) games in improving students' English speaking skills, and based on interviews and tests verified that the learning method of VRChat games improved students' English speaking skills. Literature [32] explored the effect of VR on primary school students' English performance and affective variables such as willingness to communicate and learning autonomy, based on a comparative experiment revealing that VR significantly improved students' grammar and vocabulary use in their oral performance. Literature [33] verified the positive impact of VR in improving students' language proficiency and communicative competence by examining the role of the immersionVR platform in improving students' oral proficiency, and provided insights into innovative teaching methods for English language education. Literature [34] assessed the efficacy of VR and digital storytelling (DS) technologies in improving students' English proficiency, fostering creativity and engagement, emphasizing that these technologies provide immersive experiences and creative opportunities that contribute to improving language skills. Literature [35] examined the impact of VR on the learning of English speaking skills, and a month-long study showed that VR had a positive effect on improving pronunciation, vocabulary, grammatical accuracy, and interactional skills, significantly contributing to the students' ability to express themselves in spoken English.

1.2. Purpose of the Study and Research Questions

Language acquisition is essentially a contextualized cognitive process, and the formation of English speaking ability especially depends on the immersion experience of real communication scenes. Traditional English teaching has long faced the dilemma of contextualization, which has hindered the development of learners' oral expression ability and formed the teaching problem of “mute English”. Virtual reality technology provides a technical path to solve this problem by creating a highly simulated three-dimensional interactive environment. This study aims to systematically investigate the mechanism of virtual reality English teaching on students' oral expression ability, with a special focus on the key role of multimodal data integration technology in analyzing the teaching process. This technology integration perspective breaks through the limitations of single media applications and provides a new paradigm for understanding the intrinsic mechanisms of immersive language learning.

The study focuses on three core questions: how does virtual reality technology reconfigure the cognitive context of spoken English learning? Can multimodal data fusion effectively capture the dynamic process of language acquisition? What are the key influences on learners' oral proficiency development under technological intervention? The answers to these questions need to be based on a deep understanding of the nature of language acquisition.

Ultimately, the research questions point to the optimization of teaching practices: what kind of virtual

context design can maximize the motivation for language production? How can multimodal data analysis guide personalized instructional interventions? Is there a segment-specific effect in the use of technology? The exploration of these questions will promote the formation of a new paradigm of "technology-data-teaching" for the cultivation of English speaking ability, and provide theoretical references and practical guidelines for solving the dilemma of oral language development in foreign language education.

2. Literature Review

2.1. Application of Virtual Reality Technology in Education

Virtual Reality technology brings paradigm change to the field of education through the construction of three-dimensional interactive environments, and its core value lies in the creation of an immersive learning field that breaks through the limitations of physical space. Educational neuroscience confirms that immersive environments can activate the hippocampus and prefrontal cortex synergistically, significantly improving the efficiency of knowledge encoding and extraction. This technology-driven cognitive enhancement mechanism is particularly prominent in STEM disciplines, such as in molecular biology, where learners' scores on spatial reasoning ability tests are significantly improved through dynamic manipulation of protein structure models. Exploration of virtual archaeological sites in the discipline of history has led to increased conceptual memory retention far beyond text-based learning. The technological advantage comes from the integration of multi-channel perception, the synchronization of audio-visual stimulation and somatosensory feedback to form a closed loop of embodied cognition, which transforms abstract concepts into actionable experiences.

The stimulation of learning motivation constitutes the core breakthrough of virtual reality education. Eye tracking of 8th grade students showed that visual hotspots in the virtual scene focused on contextual elements, and the average gaze duration was longer than that of PPT presentations. Motivational mechanisms can be deconstructed from self-determination theory: autonomous exploration provided by the technology satisfies the need for competence, role-playing fulfills the need for relationships, and task challenges fit the need for autonomy.

Despite the significant advantages of the technology, implementation barriers constrain its popularization. The cost of equipment poses the primary constraint, with a single VR system costing 8.3 times more than traditional multimedia equipment; the gap in teachers' technological literacy is reflected in respondents' reports of operational difficulties, reflecting the lack of a professional development system. More fundamentally, there is a risk of alienation in instructional design. A case study in a school showed that the excessive pursuit of technological showmanship overloaded the cognitive load, while the knowledge conversion rate declined. These contradictions reveal the dialectic of technology application, i.e., virtual environments need to serve pedagogical goals rather than replace pedagogical wisdom.

2.2. Research progress in multimodal data fusion techniques

The essence of multimodal data fusion technology, a cutting-edge approach in educational analytics, lies in integrating heterogeneous data sources in order to construct a holistic cognitive map of the learning process. The technique transforms multidimensional data streams such as speech, vision, and behavior into unified representation vectors through feature-level fusion strategies to form system insights beyond unimodal analysis. At the technical realization level, the synergistic architecture of Kalman filter and deep learning network constitutes the mainstream paradigm. The core of feature-level fusion lies in the introduction of the attention mechanism, which realizes the dynamic enhancement of key modalities through trainable weights.

Multimodal fusion has shown transformative potential in educational application scenarios. Empirical studies in virtual lab environments have shown that fusion models integrating eye-tracks and operation logs can predict learning bottlenecks in advance and improve intervention efficiency. This predictive advantage stems from the accurate modeling of learning state transfer by Hidden Markov Models.

The specificity of virtual reality English teaching has given rise to innovative applications of fusion technology. When learners conduct dialogue training in the immersive environment, the multimodal system synchronously collects speech spectrum, gesture space coordinates and eye movement hotspot maps, and constructs three-dimensional behavioral maps through spatio-temporal alignment algorithms. The advantage of this technology is especially significant in the error correction feedback process, where the decision model integrating speech fundamental frequency anomalies and facial confused expressions can significantly improve the acceptance rate of error correction.

The core challenge of technical evolution is the semantic gap of heterogeneous data. The non-verbal behaviors of learners in virtual environments have differences in metric scales, which need to be realized

through tensor decomposition to achieve spatial alignment of features. In the context of virtual reality English teaching and learning, fusion technology is leaping from descriptive analysis to predictive intervention. By analyzing historical interaction data, LSTM-based sequence models can predict speaking output barriers with high accuracy and trigger contextual adaptive scaffolding systems. Current research focuses on the development of transfer learning frameworks, which enable models trained in laboratory environments to be adapted to real classroom scenarios and solve the problem of skewed data distribution. This technological breakthrough will promote the formation of a closed-loop teaching paradigm of "data-driven-accurate intervention-competence development", providing a new path to solve the personalized problems in speaking teaching.

2.3. Current Research on Virtual Reality English Teaching and Learning

The research of virtual reality technology in the field of English teaching shows rapid development, and its core value is to solve the structural contradiction of the lack of context in traditional teaching. Relevant studies have confirmed that virtual contextual resources in junior high school English can significantly improve students' listening and speaking skills, and the experimental group scored significantly higher than the control group in the oral fluency test. This improvement comes from the fact that the immersive environment created by the technology activates the mirror neuron system and promotes the embodied transformation of language knowledge. Some researchers' explorations in business English teaching found that virtual negotiation scenarios led to a significant downgrade in the rate of discourse errors, confirming the effectiveness of professional scenario simulation.

Current research focuses on the analysis of the micro-mechanism of speaking ability development. A virtual tour guide system designed by a researcher confirms through scene interaction experiments that the context authenticity index is significantly positively correlated with the frequency of speaking output. Some literature found that embodied experience in elementary school AR English classes improved word memory retention, confirming the cognitive reinforcement effect of synergistic multi-sensory channels. A researcher's proposed MR deep learning methodology in upper elementary school resulted in a 2.4-fold increase in critical thinking scores, highlighting the differential impact of technology on different cognitive levels.

There are significant methodological limitations of the research paradigm. The bibliometric analysis revealed that most of the empirical studies had sample sizes of less than 200 participants, constraining the generalizability validity of the findings. At the level of instructional design, some researchers found that over-reliance on preset scripts led to a decrease in cognitive resilience, and the performance of the experimental group in open-ended speaking tasks instead showed a downward trend. Insufficient depth of technology integration constitutes a fundamental flaw, as current studies mostly use VR as a demonstration tool, failing to utilize the analytical potential of multimodal data. There are also problems such as the differentiated needs of application scenarios have not been fully satisfied, and technical implementation barriers constitute a rigid constraint for research deepening. A deeper contradiction lies in the lagging assessment system, with some researchers pointing out that current spoken language evaluation still relies on subjective scales and lacks tools to capture implicit abilities such as communicative strategies.

The research frontier is turning to the development of intelligent adaptive systems. The MR dynamic adaptation model attempted by some researchers has led to a significant increase in the progress of field-dependent students through learner characterization. Future breakthroughs need to focus on cross-scenario migration research to establish a causal chain between technological interventions and competency development, while deepening the integration of cultural dimensions in virtual contexts by design.

3. Research Methodology

3.1. Study Design

In order to systematically test the promotional effect of virtual reality English teaching on students' oral expression ability, this study adopts a mixed-methods research paradigm, integrating the technical advantages of experimental method, questionnaire survey method and content analysis method. The experimental design follows the principle of quasi-experimental research, selecting eighth grade students from two junior high schools in a city with a similar student population structure as the research subjects, and forming an experimental group (EC, n=126) and a control group (CC, n=118) through whole group sampling. There were no significant differences ($p>0.05$) between the two groups on key variables such as gender distribution, English language scores at school entry, and home language environment, ensuring that between-group comparability satisfied the premise of causal inference.

The experimental intervention lasted for 12 weeks, and 3 hours of speaking training sessions were

implemented weekly, Table 1 shows the experimental and control groups' intervention programs against each other. The experimental group was taught in a dedicated laboratory equipped with the HTC Vive Pro 2.0 system, and a self-developed virtual contextualized curriculum system was used, which included eight thematic scenarios, such as checking in at the airport, ordering food in a restaurant, and consulting a doctor in a hospital. The design of each scenario follows the principles of situational cognition theory and realizes the integration of language knowledge and communication strategies through three-dimensional spatial interaction. The control group was taught in a traditional multimedia classroom with PPT presentations and role-play activities on the same topics. The teaching was carried out by the same team of teachers, and standardized lesson plans and video analysis were used to ensure consistency of teaching and eliminate teacher effects.

Table 1. Intervention plans for the experimental group and the control group.

Dimension	Experimental Group (VR)	Control Group (Traditional)	Sig.
Presentation method of the situation	360° immersive environment	Planar image display	$p<0.001$
Interaction mechanism	Natural gesture control	Mouse click operation	$p<0.001$
Feedback system	The real-time voice analysis	Teacher's manual correction	$p<0.001$
Cognitive engagement degree	Eye-tracking value: 4.8	Eye-tracking value: 2.3	$d=1.87$
Anxiety level	Scale score: 2.7	Scale score: 5.4	$d=-1.32$

The study utilized a multidimensional assessment framework to assess oral proficiency development. The core dependent variables included fluency indicators, accuracy indicators, and communicative competence indicators, all of which were obtained from a standardized speaking test. The test was scored blindly by three certified examiners using a modified version of the Cambridge Young Learners' English Speaking Assessment System (CYLES).

A strict quality control mechanism was implemented during the study. In order to eliminate the Hawthorne effect, the control group was administered the Digital Reading Enhancement Program as a placebo intervention. The experimental group received 15 hours of training in the operation of the equipment to ensure that technical factors did not affect the effectiveness of teaching and learning. The data were analyzed using the intention-to-treat principle, and multiple interpolation was applied to those who dropped out to ensure the robustness of the results.

Ethical considerations were carried out throughout the study. Participating students signed informed consent forms, virtual scenarios were designed to avoid culturally sensitive content, and data collection was handled with anonymous coding. An additional monitoring mechanism for motion sickness was set up in the experimental group, and the intervention was immediately discontinued for those who experienced uncomfortable symptoms. The study protocol was reviewed and approved by the University Ethics Committee (Approval No. IRB-2023-EDU-015) to ensure compliance with the Declaration of Helsinki ethical guidelines for educational research.

3.2. Data collection and processing

The data acquisition system is designed with a three-tier architecture to achieve synchronized capture and real-time transmission of multimodal data through IoT technology. The sensor array deployed in the virtual reality environment contains the eye tracking module of HTC Vive Pro (sampling rate 120Hz), Shure MV7 microphone array (sampling rate 48kHz) and Leap Motion gesture recognizer (spatial resolution 0.01mm). The learning behavior data stream is precisely synchronized by a timestamp alignment algorithm, and its synchronization error is controlled within ±8ms, which meets the timing accuracy requirements of neurocognitive research. The data collection period covered the entire 12-week teaching intervention process, and a total of 3260 valid learning sessions were captured, resulting in a total of 4.7TB raw data set.

The questionnaire survey was conducted using a stratified sampling strategy, and the Virtual Learning Environment Adaptation Scale and the English Speaking Self-Efficacy Questionnaire were administered before and after the experiment. The scale design was based on the framework of situational cognition theory, containing 32 items in 5 dimensions and a 7-point Likert scale. In order to reduce the social expectation effect, the questionnaire was implemented using a double-blind process, i.e., the researchers were not involved in teaching and the teachers did not have access to the raw data. 221 valid questionnaires were recovered.

The speaking test design followed the Communicative Language Proficiency Model and consisted of three progressive tasks, i.e., Situational Response (2 minutes), Topic Elaboration (3 minutes), and Opinion Debate (5 minutes). The scoring is based on the modified IELTS Speaking Scale, which is a 9-point scale evaluating four dimensions: fluency, accuracy, vocabulary and pragmatic ability. The team of examiners consisted of four TESOL-certified teachers with 20 hours of standardized training and an intra-group correlation coefficient of 0.87 ($p<0.001$). The entire testing process was videotaped and transcribed into a text corpus for subsequent analysis.

The multimodal data were processed using a feature-level fusion strategy with a mathematical model expressed as:

$$F(x) = \alpha_1 \cdot \phi(V) + \alpha_2 \cdot \psi(E) + \alpha_3 \cdot \gamma(G) + \alpha_4 \cdot \eta(F) \qquad (1)$$

where V represents the speech feature vector, E is the eye movement feature matrix, G is the gesture movement tensor, and F represents the facial expression coding. The feature mapping functions ϕ, ψ, γ, η are realized as dimensionality reduction using a deep convolutional self-encoder to transform the original data into a 128-dimensional latent space representation, respectively. The weight coefficients α_i are dynamically computed by the entropy weighting method, reflecting the information contribution of each modality at a specific learning stage.

The data processing process contains four key stages: the original signal is eliminated from device noise by a Butterworth filter; feature extraction stage adopts sliding window analysis; modal alignment realizes time-domain matching by a dynamic time regularization algorithm; and the fusion calculation adopts tensor decomposition technique. Namely:

$$\mathrm{T} = \sum_{r=1}^{R} \lambda_r \cdot u_r \circ v_r \circ w_r \qquad (2)$$

where T is the third-order fusion tensor, the model successfully captures cross-modal correlations ignored by traditional methods, such as the synergistic patterns of gesture amplitude and speech pauses.

The qualitative data were processed using the Rooted Theory analytic framework with three levels of coding for the 48 transcribed teacher-student interviews. Open coding identified the core categories of "situational authenticity" (frequency 142) and "immediacy of feedback" (frequency 118); Axial coding establishes a causal chain of "technology acceptance→cognitive load→learning effectiveness"; selective coding forms a theoretical model of "embodiment-contextualization-feedback", and the theoretical saturation test is reached after the seventh iteration. To enhance the analytic validity, the Nvivo software was used to implement the coder consistency test (Kappa = 0.82), and the interpretive validity was confirmed by the respondents through the member-checking method.

The multimodal data warehouse was constructed using a star architecture, where the fact table records the 128-dimensional fusion features of each learning session, and the dimension table contains metadata such as learner attributes and contextual parameters, and this design supports complex OLAP queries. The data visualization subsystem uses D3.js to develop interactive dashboards that present individual learning trajectories and group distribution characteristics in real time, providing decision support for instructional interventions.

3.3. Analysis of Research Reliability

The research reliability and validity test adopts a multi-level validation framework to ensure the scientificity of the data and the reliability of the conclusions through the strategy of combining quantitative analysis and qualitative testing, and the results of the analysis of the core indicators of reliability and validity are shown in Table 2. The internal consistency test and validation factor analysis were used in the questionnaire stage, and the Cronbach's alpha coefficient of the Virtual Learning Environment Adaptation Scale reached 0.89, which exceeded the benchmark requirement of 0.7 in psychometrics. Validated factor analysis showed excellent model fit indicators ($\chi^2 / df = 1.87$, CFI=0.93, RMSEA=0.048), confirming that the structural validity of the scale meets the research needs. The retest reliability of the scale maintained a high stability of 0.82 after a two-week interval, reflecting the temporal consistency of the measurement instrument. The validity scale correlation validity test was analyzed by Pearson correlation with the Technology Acceptance Questionnaire, indicating that the scale was effective in capturing the process of psychological adaptation to technology interventions.

The reliability of the speaking test was assured using a three-stage control mechanism. Inter-rater reliability was quantified by the intragroup correlation coefficient (ICC), and the agreement of the three examiners' scores on the impromptu speaking task reached 0.86 (95% CI:0.82-0.90), which satisfies the stringent criterion of 0.75 in clinical psychology. The content validity test was performed by five applied

linguistics experts, and three rounds of correction were applied to the test tasks using the Delphi method, resulting in a final content validity index (CVI) of 0.93. The structural validity of the test was verified by a multi-trait, multi-method matrix, and the convergent validity of the fluency and accuracy dimensions (r=0.68) was significantly higher than the discriminant validity (r=0.21), which confirms the reasonableness of the measurement dimensionality delineation. Parallel test reliability showed a correlation of 0.84 in A/B paper tests taken one week apart, excluding systematic errors due to test format.

Table 2. Analysis of core indicators of reliability and validity.

Data category	Inspection dimension	Measure index	Result	Standard
Questionnaire survey	Internal consistency	Cronbach's α	0.89	>0.70
Questionnaire survey	Structural validity	CFI/RMSEA	0.93/0.048	>0.90/ <0.08
Oral test	Rater reliability	ICC	0.86	>0.75
Oral test	Content validity	CVI	0.93	>0.80
Eye movement data	Equipment accuracy	Spatial resolution	0.48°	<1.0°
Voice data	Sampling quality	Signal-to-noise ratio	32.7dB	>25dB
Interview text	Coding consistency	Kappa coefficient	0.83	>0.75

The confidence challenge of multimodal data mainly stems from the device synchronization error and feature extraction variance. To address the problem of time asynchrony, the developed time-domain alignment algorithm controls the eye-movement-speech data synchronization error within ±8ms, which is 63% higher than the traditional method. Feature stability is verified by calculating the intra-group correlation coefficient, and the gesture motion entropy has an ICC of 0.86 in repeated tests, confirming the cross-situational consistency of behavioral features. The system error calibration adopts the standard reference method, and the baseline articulator model is set in the virtual scene, and the error rate of speech feature extraction is reduced to 3.2%. The control variables of the data acquisition environment include light intensity (maintained at 500 ± 50 Lux) and background noise (<35 dB) to eliminate the measurement variance caused by environmental interference.

4. Findings and Analysis

4.1. Analysis of the results of the questionnaire

The questionnaire data reveal the deep impact of virtual reality English teaching on students' psycho-cognitive structure, and its specific results are shown in Table 3. As can be seen from the table, the mean score of students in the experimental group in the dimension of learning interest amounted to 4.32 (SD=0.56), which was significantly higher than that of the control group, which was 3.45 (SD=0.61), and the t-test of independent samples showed that the difference between the groups reached a statistically significant level (t=7.83, p<0.001, d=1.42). This increased interest showed a cumulative effect over time, with the mean interest score of 3.98 at week 4 of the intervention increasing steadily to 4.32 by week 12, and the goodness-of-fit of the linear regression model (0.89) suggesting a sustained motivational effect of the technology intervention. The change curve of motivation intensity was more enlightening, as the intrinsic motivation score of the experimental group jumped from baseline 3.05 to 4.25 (Δ=39.34%), while the control group only increased by 9.6%, confirming the reinforcing mechanism of the virtual context on the fulfillment of competence needs in the self-determination theory.

Contextual engagement data reveals the immersive effect of virtual environments. Students in the experimental group reported a cognitive engagement score of 4.56 in the Airport Check-In scenario, significantly higher than the 3.32 in the traditional classroom, and the difference in engagement was due to the synergistic activation of multi-sensory channels, with 78.3% of the experimental group reporting that they “forgot that they were in the classroom” compared to 12.4% of the control group. 78.3% of students in the experimental group reported “completely forgetting that they were in the classroom,” compared to 12.4% in the control group. Path analysis revealed a standardized effect coefficient of 0.67 (p<0.001) for situational realism on learning interest, confirming the applicability of situated cognition theory to virtual language learning. It is worth noting that there is a gender difference in the increase in engagement, with boys' engagement in the virtual environment increasing significantly more (46.7%) than girls' (32.1%), reflecting the differentiated impact of technological interventions on different learning styles.

The analysis of technology acceptance was characterized by a U-shaped curve. The mean technology anxiety score of the experimental group reached 3.85 at the beginning of the intervention (1-2

weeks), which mainly originated from the complexity of equipment operation. It dropped rapidly to 2.12 by week 4, reflecting the steep characteristics of the adaptive learning curve. The technical affinity score soared to 4.56 at week 12, creating a significant shift in technical identity. Structural equation modeling verified the direct effect of perceived ease of use on attitude toward use ($p<0.001$), while perceived usefulness mediated the effect on willingness to continue use through satisfaction ($p<0.001$).

Table 3. Psychological impact questionnaire survey.

Dimension	Group	Before test	After test	t	Effect size
Learning interest	EC	3.21	4.32	7.83***	*d*=1.42
	CC	3.18	3.45		
Intrinsic motivation	EC	3.05	4.25	10.27***	*d*=1.92
	CC	3.02	3.31		
Situational engagement degree	EC	2.87	4.18	11.36***	*d*=2.15
	CC	2.91	3.22		
Technical acceptance	EC	3.32	4.56	12.18***	*d*=2.31
	CC	3.35	3.41		

4.2. Analysis of the results of the speaking test

The quantitative results of the speaking test revealed the structured impact of VRT on language output, the comparative results of which are shown in Table 4. Students in the experimental group achieved a posttest mean of 7.32 in the fluency dimension, an improvement of 42.69% over the pre-test, a gain significantly higher than that of the control group, which was 15.35% ($t=9.47, p<0.001, d=1.83$). This improvement was particularly prominent in the impromptu conversation task, where the experimental group's effective vocabulary per minute increased from 98.5 to 156.2, while the control group's increased only to 112.3. The nonlinear growth feature of fluency was revealed in the segmented regression model, where the experimental group showed an inflection point at week □6, and the rate of improvement jumped from 0.12 to 0.31 points per week, reflecting the catalytic effect of the immersive environment on language automation. The improvement in phonological accuracy showed modality specificity. Consonant cluster articulation error rate decreased to 11.4% in the experimental group, which was significantly better than 26.7% in the control group ($\chi2=18.33, p<0.001$). Acoustic analysis showed a 32.5% improvement in phoneme boundary clarity in the virtual group, with resonance peak migration trajectories closer to the target language model. This improvement was strongly correlated with eye-movement data, with articulatory accuracy up to 89.2% when the virtual character's lips stayed in the center of the visual field for more than 800ms, confirming the value of visual cues in guiding speech imitation. It is worth noting that the improvement of intonation patterns by technological intervention was limited, with the experimental group's realization rate of rising intonation in interrogative sentences increasing by only 12.3%, reflecting the need for a finer auditory feedback mechanism for prosodic acquisition. The improvement in pragmatic competence reflects the advantage of contextual adaptation. In the restaurant ordering situation, the experimental group's pragmatic appropriateness score reached 6.77, 24.7% higher than that of the control group. The multimodal fusion data reveals that this improvement stems from the synergistic optimization of gesture and speech. When asked by the virtual environment to hand over an object, 83.6% of the students in the experimental group synchronized their gestures accordingly, while only 37.2% of the control group achieved action-verbal integration. This embodied expression resulted in a 41.3% increase in communicative efficiency and a reduction in turn-taking latency from 2.3 to 1.1 seconds. The between-group difference in the rate of cultural pragmatic errors was even more convincing, as the experimental group's error frequency in the cross-cultural scenario was 0.8 errors/minute, which was significantly lower than that of the control group's 2.4 ($p<0.001$), confirming the facilitating effect of the virtual situation on the internalization of cultural scripts.

Table 4 Comparison of the performance of the divided dimension.

Dimension	Group	Before test	After test	t	Effect size
Learning interest	EC	5.13(0.82)	7.32(0.78)	9.47***	*d*=1.83
	CC	5.08(0.79)	5.86(0.81)		
Intrinsic motivation	EC	4.87(0.91)	6.95(0.85)	14.87***	*d*=1.36

	CC	4.92(0.88)	5.61(0.87)		
Situational engagement degree	EC	4.75(0.95)	6.83(0.92)	13.25***	*d*=2.47
	CC	4.71(0.93)	5.54(0.89)		
Technical acceptance	EC	4.68(1.02)	6.77(0.96)	12.98***	*d*=2.15
	CC	4.65(0.98)	5.43(0.94)		

4.3. Analysis of interview results

Deep mining of teacher and student interview data reveals a multidimensional picture of the virtual reality English teaching experience. The content analysis method identifies four core categories, namely, contextual immersion experience, perception of technological interaction, psychological transformation of learning and pedagogical adaptation challenges. Table 5 shows the analysis matrix of the core categories of the teacher-student interviews. Contextual immersion experience showed significant scene differences, with the airport check-in scene scoring significantly higher than the hospital scene in terms of sense of presence ($t=5.37, p<0.001$). This difference stems from the coupling effect of spatial complexity and interaction density, and the number of interactable objects in the 3D environment is strongly correlated with immersion ($r=0.78, p<0.01$). Typical narratives such as "When the virtual customs officer looked directly at me and asked me a question, I unconsciously answered it with a complete sentence, which is completely different from memorizing a conversation in the classroom" confirm the practical value of embodied cognition theory in virtual language learning.

Perceptions of technological interactions showed a double-edged sword effect. Positive experiences focused on natural interaction mechanisms, with 82.7% of students believing that "gesture control is more in line with real conversation habits than mouse clicks". Eye-tracking data supports this view, with the experimental group's visual-motor coordination time in the VR environment reduced to 0.87s, a 63% improvement over traditional interfaces. However, device suitability poses a major obstacle, with 41.5% of teachers reporting that "the weight of the helmet makes it inadvisable to train for more than 25 minutes in a single session" and a 34.2% incidence of vertigo in field-dependent students. This contradiction is materialized in the following narrative, "Although it was a bit boring to wear the helmet, I couldn't help ordering in English when the virtual waiter smiled at me" (experimental group), reflecting the psychological game between technological barriers and contextual attraction.

Table 5. Interview core category analysis matrix.

Core Category	Typical narrative	Frequency	Emotion	Ratio
Immersive situational experience	The shop signs on the virtual street are all in English, giving a real feeling of being abroad	142	Positive (93.7%)	38.6%
Technology interactive perception	Gesture recognition is sometimes delayed, interrupting the smoothness of the conversation	118	Negative (67.8%)	6.4%
Learning psychological transformation	When I make a mistake, the virtual character will encourage me. Now I dare to speak up voluntarily	105	Positive (89.5%)	25.3%
Teaching adaptation challenge	Additional VR-specific teaching plans need to be prepared, doubling the workload	76	Negative (82.9%)	29.7%

The psychological transformation of learning was characterized by stages. At the beginning of the intervention, the anxiety index amounted to 4.15 (out of 5 on the scale), mainly stemming from technological unfamiliarity; the self-confidence score jumped to 4.32 after week 4, which was significantly correlated with the improvement in fluency ($t=0.71, p<0.001$). This shift was particularly significant among socially anxious students, whose frequency of classroom presentations increased from 0.5/week to 4.2/week. Teacher observation noted that "students who usually kept their heads down and did not speak, took the initiative to recommend dishes to virtual customers in the VR restaurant" (experimental group). The analysis of psychological mechanisms reveals that the "error-tolerant feedback" of virtual characters is the key, and 78.3% of students believe that "virtual listeners will not laugh at mispronunciation", which lowers the emotional filtering barrier. The challenge of teaching adaptation focuses on the problem of resource allocation. The equipment sharing rate is as high as 5.2 students per unit, resulting in less than 18 minutes of training in a single day. There is a professional disconnect in school-based curriculum development, with only 32.4% of teachers able to modify virtual scenarios on their own. A more prominent contradiction is the mismatch in the evaluation system, with

traditional paper-and-pencil tests failing to capture the progress of discourse competence in the VR environment. Interviews with principals pointed out that “the existing assessment indicators focus on grammatical accuracy, but the resilience that students show in virtual customs conversations is more worthy of recognition” (experimental school). This structural contradiction constrains the maximization of technological benefits.

5. Conclusion

The facilitating effect of virtual reality English teaching on students' oral expression ability was confirmed in several dimensions. Students in the experimental group significantly improved their posttest scores in the oral fluency dimension by 42.7% compared to the control group. This improvement was particularly prominent in the impromptu conversation task, where effective vocabulary per minute significantly exceeded that of the traditional teaching group. Improvements in phonological accuracy showed modality specificity, with acoustic analyses revealing increased clarity of phoneme boundaries, and this improvement was strongly correlated with eye movement data. The improvement in pragmatic competence reflects the advantage of contextual adaptation, and the frequency of pragmatic errors in cross-cultural scenarios is only 0.8 per minute, significantly lower than the 2.4 in the control group, confirming the facilitating effect of the virtual context on the internalization of cultural scripts. The application of multimodal data fusion technology creates a new paradigm for learning analysis, and learning motivation and contextual immersion constitute the psychological basis for the realization of the effect.

Funding

Granting Organization: Supply-Demand Docking Employment and Education Program of the Ministry of Education: Title: Research on Innovative Strategies for the Cultivation Model of Targeted English Talents in Universities under the Background of Digital Transformation (Year: 2024) (2024122570446).

About the Author

Qijun Zhao, gender: female, 1983.10, Yi, Ludian County, Yunnan Province; Education: Master's degree; Title: Associate Professor; Research interests: Applied Linguistics and English Teaching Methodology.

References

1. Chapple, J. (2015). Teaching in English is not necessarily the teaching of English. International Education Studies, 8(3), 1-13.
2. Kuchah, K. (2018). Teaching English in difficult circumstances: Setting the scene. In International perspectives on teaching English in difficult circumstances: Contexts, challenges and possibilities (pp. 1-25). London: Palgrave Macmillan UK.
3. Shan, L. W., & Aziz, A. A. (2022). A systematic review of teaching English in rural settings: Challenges and solutions. International Journal of Academic Research in Business and Social Sciences, 12(6), 1956-1977.
4. Agung, A. S. N. (2019). Current challenges in teaching English in least-developed region in Indonesia. SOSHUM: Jurnal Sosial dan Humaniora, 9(3), 266-271.
5. Abrar, M. (2016, October). Teaching English problems: An analysis of EFL primary school teachers in Kuala Tungkal. In The 16th Indonesian Scholars International Convention (pp. 94-101).
6. Marzulina, L. (2021). Challenges in Teaching English. Theory and Practice in Language Studies, 11(12), 1581-1589.
7. Bahari, A. (2022). Affordances and challenges of teaching language skills by virtual reality: A systematic review (2010–2020). E-Learning and Digital Media, 19(2), 163-188.
8. Reitz, L., Sohny, A., & Lochmann, G. (2016). VR-based gamification of communication training and oral examination in a second language. International Journal of Game-Based Learning (IJGBL), 6(2), 46-61.
9. Luo, X. (2022). Practice of artificial intelligence and virtual reality technology in college English dialogue scene simulation. Wireless Communications and Mobile Computing, 2022(1), 4922675.
10. Chiew, F. L., Kho, G. H. X., Kong, M. Y., Lu, Y. Z., Yunus, M. M., Hashim, H., & Syafril, S. (2025). Augmented Reality (AR) in ESL Classrooms: A Quasi-Experimental Study on Enhancing Speaking Skills. Karya Journal of Emerging Technologies in Human Services, 1(1), 38-47.
11. Yang, G., Chen, Y. T., Zheng, X. L., & Hwang, G. J. (2021). From experiencing to expressing: A virtual reality approach to facilitating pupils’ descriptive paper writing performance and learning behavior engagement. British Journal of Educational Technology, 52(2), 807-823.
12. Shah, D. S. M., Othman, S., Salim, M. S. A. M., Salim, M. N. F. M., Khalil, M. I. M., & Kusmawan, U. (2024). The impact of immersive 360-degree video learning on enhancing oral communication skills. Journal of Advanced Research in Applied Sciences and Engineering Technology, 58(1), 55-71.
13. Akram, H., Yang, Y., Ahmad, N., & Aslam, S. (2020). Factors contributing low English language literacy in rural primary schools of Karachi, Pakistan. International Journal of English Linguistics, 10(6), 335-346.

14. Oeamoum, N., & Sriwichai, C. (2020). Problems and Needs in English Language Teaching from the Viewpoints of Preservice English Teachers in Thailand. Asian Journal of Education and Training, 6(4), 592-601.
15. Al-Seghayer, K. (2014). The four most common constraints affecting English teaching in Saudi Arabia. International Journal of English Linguistics, 4(5), 17.
16. Abrar, M., Mukminin, A., Habibi, A., Asyrafi, F., Makmur, M., & Marzulina, L. (2018). If our English isn'ta language, what is it? Indonesian EFL student teachers' challenges speaking English. The Qualitative Report, 23(1), 129-145.
17. Anyiendah, M. S. (2017, May). Challenges faced by teachers when teaching English in public primary schools in Kenya. In Frontiers in Education (Vol. 2, p. 13). Frontiers Media SA.
18. Yang, F. C. O., Lo, F. Y. R., Hsieh, J. C., & Wu, W. C. V. (2020). Facilitating communicative ability of EFL learners via high-immersion virtual reality. Journal of Educational Technology & Society, 23(1), 30-49.
19. Chang, H., Park, J., & Suh, J. (2024). Virtual reality as a pedagogical tool: An experimental study of English learner in lower elementary grades. Education and Information Technologies, 29(4), 4809-4842.
20. Li, J., Qiu, T., Li, C., Xu, C., Cheng, P., Tang, Y., & Georgiou, G. K. (2024). The effects of immersion in virtual reality environment on oral English learning for Chinese university students. Education and Lifelong Development Research, 1(1), 3-14.
21. Li, X., Xie, Y., & Liu, T. (2020, May). Research on oral English teaching system based on vr in the background of ai. In Journal of Physics: Conference Series (Vol. 1550, No. 2, p. 022031). IOP Publishing.
22. Lou, Y. (2025). The impact of virtual reality environments on English language acquisition: Innovative immersive learning technologies for communication skills development. Journal of Computational Methods in Sciences and Engineering, 14727978251337950.
23. Ahmet, A. C. A. R., & Cavas, B. (2020). THE EFFECT OF VIRTUAL REALITY ENHANCED LEARNING ENVIRONMENT ON THE 7TH-GRADE STUDENTS'READING AND WRITING SKILLS IN ENGLISH. MOJES: Malaysian Online Journal of Educational Sciences, 8(4), 22-33.
24. Alshumaimeri, Y. A., & Alhumud, A. M. (2021). EFL Students' Perceptions of the Effectiveness of Virtual Classrooms in Enhancing Communication Skills. English Language Teaching, 14(11), 80-96.
25. Saeedizadeh, E., & Khodabandeh, F. (2024). Unleashing the Power of Engage Virtual Reality Learning App: A Syntactic Complexity Boost in Intermediate EFL Students' Oral Performance. PRESENCE: Virtual and Augmented Reality, 33, 315-337.
26. Yang, Q. (2025). Training of English Listening and Speaking Using Virtual Reality Technology. Journal of the Brazilian Computer Society, 31(1), 711-716.
27. Yudintseva, A. (2023). Virtual reality affordances for oral communication in English as a second language classroom: A literature review. Computers & Education: X Reality, 2, 100018.
28. Park, H. (2022). Effects of Virtual Reality-based English Learning on Korean University Students' Speaking Ability. Multimedia-Assisted Language Learning, 25(4).
29. Ebadi, S., & Ebadijalal, M. (2022). The effect of Google Expeditions virtual reality on EFL learners' willingness to communicate and oral proficiency. Computer Assisted Language Learning, 35(8), 1975-2000.
30. Muhammad, R. (2023). Enhancing English Oral Skills among Malaysian Rural School Students through the Implementation of Virtual Reality (VR). International Journal on E-Learning Practices (IJELP), 6(1).
31. Cahyadi, P., Wardhana, D. I. A., Ansori, W. I., & Farah, R. R. (2022). ENHANCING STUDENTS'ENGLISH SPEAKING ABILITY THROUGH VRCHAT GAME AS LEARNING MEDIA. Journal of Research on Language Education, 3(2), 54-61.
32. Sally Wu, Y. H., & Alan Hung, S. T. (2022). The effects of virtual reality infused instruction on elementary school students' English-speaking performance, willingness to communicate, and learning autonomy. Journal of Educational Computing Research, 60(6), 1558-1587.
33. Raman, K., Hashim, H., & Ismail, H. H. (2024). Exploring the Impact of VR Integration on ESL Leaners' English Verbal Communication Skills: A Case Study in a Malaysian High School. Arab World English Journal.
34. Azir, I. D. A., Sriyanto, W., Sitorus, N., & Anggria, F. (2024). Virtual reality (VR) and digital storytelling (DS) technology to improve English speaking skills of vocational students. KnE Engineering, 295-305.
35. JUMASHOVA, Z. (2024). EVALUATING VIRTUAL REALITY AS A TOOL FOR IMPROVING SPEAKING SKILLS IN ENGLISH LANGUAGE LEARNING. PedActa, 14(2).