Region-Educational Research and Reviews, 2025,Volume 7, Issue 8
DOI:10. 32629/rerr.v7i8.4512
ISSN Online:2661-4634
ISSN Print:2661-4626

Research on the Integration of English Listening and Speaking with Artificial Intelligence from an Interdisciplinary Perspective

Jie Chi, Wei Yang

Nanjing University of Science and Technology, Nanjing, Jiangsu, 210094, China

Abstract: This study explores the deep integration of English listening and speaking instruction with artificial intelligence technology from an interdisciplinary perspective. By analyzing the neural mechanisms underlying speech cognition, we developed a deep learning-based intelligent teaching system architecture. The research employs core technologies such as automatic speech recognition and neural network speech synthesis to design personalized training models. Empirical evidence demonstrates that AI-assisted teaching significantly outperforms traditional methods in enhancing learners' speech perception accuracy and oral fluency. The findings provide theoretical foundations and technical pathways for the digital transformation of foreign language education, while promoting interdisciplinary innovation in computational linguistics and second language acquisition theory.

Keywords: artificial intelligence, English listening and speaking, interdisciplinary, speech recognition, intelligent teaching, second language acquisition

1. Introduction

With groundbreaking advancements in deep learning technology, artificial intelligence has demonstrated tremendous potential in language education. As core communicative skills in English language learning, the effectiveness of listening and speaking instruction directly impacts learners' language proficiency. Traditional teaching models face limitations in personalized guidance, real-time feedback, and large-scale implementation. Grounded in interdisciplinary theories from neuroscience, cognitive psychology, and computational linguistics, this study aims to develop an intelligent English listening and speaking teaching system. By simulating human speech cognition through technological means, it achieves precise speech analysis, intelligent dialogue interaction, and personalized learning path planning. This cross-disciplinary research not only provides innovative tools for foreign language education but also pioneers new directions for theoretical frameworks and practical applications of human-machine collaborative learning models.

2. Cognitive Mechanisms of English Listening and Speaking Skills with AI Simulation

2.1 Neuroscience Basis of Speech Perception and Production

Human speech perception and production involve complex neural network coordination mechanisms. The auditory cortex processes acoustic signals into linguistic information through hierarchical processing, where the superior temporal gyrus extracts fundamental acoustic features while the middle temporal gyrus participates in temporal integration of

speech sequences. Language production primarily relies on the coordinated interaction between the left hemisphere's Broca's area and Wernicke's area—the former controls speech motor planning, while the latter handles semantic selection and syntactic organization. Research reveals significant differences in speech processing networks between second language learners and native speakers, manifested as increased right hemisphere involvement and excessive activation in frontal lobe regions. These neuroplasticity characteristics provide a biological basis for personalized language training, while simultaneously revealing critical period effects and cognitive load mechanisms in speech learning.

2.2 Modeling of Human Auditory-Cognitive Processes by Artificial Intelligence

Deep neural networks provide effective tools for simulating human speech cognition. Convolutional neural networks mimic the hierarchical processing mechanism of the auditory cortex through multi-layer feature extraction, while recurrent neural networks capture temporal dependencies in speech sequences. The introduction of attention mechanisms enables models to dynamically focus on critical speech segments, mirroring human selective auditory attention processes. The Transformer architecture further achieves parallel processing and long-distance dependency modeling, effectively addressing the gradient vanishing problem in traditional sequence models. In speech generation modeling, variational autoencoders and generative adversarial networks (GANs) learn latent representation spaces of speech, enabling end-to-end mapping from semantics to acoustics. Multimodal fusion technology integrates speech, text, and visual information, constructing a more comprehensive communicative competence model.

3. Technical Architecture and Implementation of the Intelligent English Listening and Speaking Teaching System

3.1 Application of automatic speech recognition technology in hearing training

Modern speech recognition systems employ end-to-end deep learning architectures that directly map audio signals into text sequences through acoustic models. To address the speech characteristics of non-native learners, the system utilizes domain adaptation techniques and data augmentation methods to optimize acoustic modeling, enhancing robustness against accent variations and pronunciation errors. The real-time evaluation module, powered by confidence analysis and speech quality detection algorithms, precisely identifies listening comprehension challenges for learners. A personalized difficulty adjustment mechanism dynamically modifies speech materials 'speed, complexity, and noise levels based on learners' historical performance and cognitive load levels [1]. Additionally, the system incorporates rhythm analysis capabilities to help learners master English stress patterns, intonation, and rhythmic structures, thereby comprehensively improving listening comprehension accuracy and efficiency.

3.2 Teaching function development of speech synthesis and dialogue system

Neural network speech synthesis technology has significantly enhanced the naturalness and expressiveness of synthesized speech. Models like WaveNet and Tacotron can generate speech approaching human-level quality, providing high-quality demonstration materials for oral practice. Multi-speaker speech synthesis systems support voice generation across different genders, ages, and accents, enriching learners' speech input environments. Intelligent dialogue agents, built on large-scale pre-trained language models, possess contextual understanding, emotion recognition, and pragmatic reasoning capabilities, enabling natural and fluent English conversations. The system automatically generates diverse communication scenarios including academic discussions, business negotiations, and casual chats, offering learners abundant opportunities for oral practice. The dialogue management module adjusts topic difficulty and interaction strategies based on learning objectives, ensuring targeted and effective practice content. In terms of speech synthesis detail control, the system achieves fine-grained regulation of rhythmic features including pitch contour, duration distribution, and energy variations, making synthesized speech more realistic for communicative contexts. The combined architecture of FastSpeech2 and HiFi-GAN ensures dual optimization of synthesis speed and audio quality. The dialogue system adopts a hybrid architecture combining retrieval-based and generative approaches, ensuring both accuracy in

responses and enhanced expressive flexibility.

3.3 Multimodal Interface Design and User Experience Optimization

The multimodal interaction system integrates voice, gestures, eye movements, and facial expressions to create an immersive language learning environment. Virtual reality technology constructs realistic English usage scenarios such as airports, restaurants, and meeting rooms, enhancing learners 'presence and engagement. The affective computing module monitors learners' emotional states in real-time through voice emotion recognition and facial expression analysis, promptly adjusting teaching strategies and interaction methods. The adaptive interface presents personalized content based on learners' operational habits and cognitive styles, optimizing information layout and interaction flow. Gamification elements like point systems, achievement badges, and leaderboards stimulate intrinsic motivation. The system also provides detailed learning analytics reports to help learners track progress, identify weaknesses, and develop more precise study plans[2].

4. Theoretical innovation and practical effect evaluation of interdisciplinary integration

4.1 The intersection of computational linguistics and second language acquisition theory

The integration of interdisciplinary theories has introduced innovative research paradigms and practical models for English listening and speaking instruction. Through large-scale learner corpus analysis, the system identifies common patterns in second language development and individual differences, providing data-driven validation for Selinker's Second Language Acquisition (SLA) theory. The Error Analysis Theory is algorithmically implemented in intelligent error correction systems, where deep learning models automatically classify speech errors—including phonological substitutions, prosodic deviations, and intonation anomalies—and provide corresponding corrective strategies. Krashen's Input Hypothesis is technologically adapted through comprehensible input algorithms, dynamically adjusting vocabulary complexity and syntactic structures based on learners' current proficiency levels to ensure input content remains within the optimal acquisition zone. Guided by communicative teaching principles, the system emphasizes meaning negotiation and authentic contextualization, ensuring technological tools serve the fundamental goal of developing communicative competence.

4.2 Effect Verification and Future Development of Intelligent Listening and Speaking Teaching

An empirical study employed a quasi-experimental design to compare the effectiveness of traditional teaching methods versus AI-assisted instruction. Sixteen English learners were randomly assigned to an experimental group and a control group for an 8-week teaching experiment. Results demonstrated that the experimental group significantly outperformed the control group in speech recognition accuracy, oral fluency, and communicative willingness, with effect sizes reaching 0.78,0.65, and 0.52 respectively. The multidimensional evaluation system encompassed metrics such as pronunciation accuracy, natural rhythm, lexical richness, grammatical complexity, and communicative outcomes. By combining machine scoring with human evaluation, the study ensured objectivity and comprehensiveness. Longitudinal data revealed nonlinear learning trajectories and individual variation patterns, providing empirical evidence for optimizing personalized teaching approaches [3]. Future development should balance technological innovation with humanistic care, fully leveraging AI's technical advantages while maintaining the human warmth of language education and the irreplaceable value of teacher-student emotional interaction.

Conclusion

This study explores interdisciplinary pathways for integrating English listening/speaking instruction with artificial intelligence (AI) technology, achieving significant advancements in both theoretical frameworks and practical applications. The research demonstrates that neural mechanisms underlying speech cognition provide biological inspiration for AI model design, while deep learning technologies offer robust tools for language education. The

developed intelligent teaching system enables personalized learning, real-time feedback, and immersive interactions, significantly enhancing learning outcomes and user experience. Interdisciplinary theoretical integration fosters synergistic development between computational linguistics and second language acquisition research, laying a solid foundation for the digital transformation of foreign language education. However, critical issues such as data privacy protection, algorithmic fairness, and educational ethics remain to be addressed in technology implementation. Future research should further explore optimal human-machine collaboration models, maintaining the humanistic essence of education while leveraging technological advantages to promote sustainable development of language education in the intelligent era.

References：

[1] Wensheng Chen, Shi Zhaohua. How Artificial Intelligence is Reshaping Intelligence Studies: New Explorations from an Interdisciplinary Perspective [J]. Intelligence Exploration, 2025(1):19-28.

[2] Xinning Huang. Integrating Artificial Intelligence into Foreign Language Teaching at Universities: A Cross-disciplinary Perspective [J]. Journal of Taiyuan Vocational and Technical College, 2024(11):86-88.

[3] Jianlin Bai, Zezhi Zhang, Wenwen Sun. A Study on Cultivation Strategies for Key AI Teachers from an Interdisciplinary Perspective [J]. Primary and Secondary School Information Technology Education, 2023(8):43-44.