The Influence of Intelligent Speech Recognition on the Interactivity and Learning Efficiency of Vocal Music Teaching

Kunming Wang[1,*], Ying Lei[2]
[1]School of Training and Continuing Education, Guangdong Polytechnic of Industry and Commerce, Guangzhou, China
[2]School of Applied Foreign Languages, Guangdong Polytechnic of Industry and Commerce, Guangzhou, China
** Corresponding Author*

Abstract: This study aims to explore the application effect of intelligent speech recognition in vocal music teaching. A quasi-experimental design was adopted, with 60 undergraduate students as the subjects. The experimental group was introduced with intelligent speech recognition for real-time analysis and feedback, while the control group continued with traditional teaching. Data were collected through classroom observations, vocal tests, questionnaires and interviews, and analyzed using statistical software. The results showed that the experimental group was significantly superior to the control group in terms of the frequency of classroom speeches, the number of interaction rounds and the quality of teachers' feedback. In terms of learning efficiency, the experimental group performed better in pitch accuracy, rhythm, overall singing quality and task completion speed. Further analysis indicates that cognitive load plays a partial mediating role between teaching mode and learning efficiency, and technological acceptance has a moderating effect on teaching effectiveness. The conclusion holds that intelligent speech recognition can effectively enhance the interactivity and learning efficiency of vocal music classes, but attention should be paid to the stability of the technology and the dependence of learners. Future research can further verify its educational potential in larger samples and multi-dimensional scenarios.

Keywords: Intelligent Speech Recognition; Vocal Music Teaching; Classroom Interactivity; Learning Efficiency; Educational Technology

1. Introduction

In recent years, the rapid development of artificial intelligence (AI) technology has driven profound changes in educational methods, especially achieving remarkable progress in personalized learning, real-time feedback and interactive teaching. Merchán Sánchez-Jara et al. [1] pointed out through a systematic literature review that AI in music education is developing in a more personalized, interactive and efficient direction, covering multiple fields such as intelligent tutoring systems, interactive auditory training, and assessment systems. Meanwhile, Zhang's research also focuses on applying big data and AI to online vocal music smart classrooms to enhance the personalization and effectiveness of teaching [2]. These studies have laid the foundation for the integration of intelligent technology and traditional art teaching, and also provided theoretical and empirical support for this research to explore the role of intelligent speech recognition in vocal music teaching.

Intelligent speech recognition, as an important branch of AI, has been widely applied in language teaching and educational interaction. For instance, AI voice assistants can optimize students' learning behaviors and the ways of interaction between teachers and students in educational contexts, enhancing the responsiveness and efficiency of teaching [3]; In primary school English listening and speaking teaching, intelligent speech recognition technology significantly promotes the improvement of students' language ability through immediate feedback [4]. In addition, artificial intelligence is widely applied in scenarios such as personalized tutoring and emotional companionship in children's education, which helps improve learning quality. However, it also brings challenges such as ethical privacy

and educational equity [5]. In the field of foreign language education, the participation of AI models helps enhance students' autonomy and efficiency, but excessive reliance may harm the development of basic abilities or the deepening of cultural connotations [6]. The research results from these different educational backgrounds suggest that introducing intelligent speech recognition technology into the interaction of vocal music teaching may not only deepen the dynamic feedback between teachers and students, but also pay attention to the preservation of students' subjectivity and cultural expression.

In terms of the specific application of vocal music teaching, existing research has initially verified the feasibility and advantages of speech recognition and AI technology. In a study published in "Artificial Intelligence in Music Education" in 2024, by using pre-trained speech recognition models and deep learning methods, the vocal line characteristics of vocal singers were analyzed. The results show that AI-assisted training can effectively distinguish the performance differences between the experimental group and the control group of learners [7]. Furthermore, Jin et al. [8] studied the application of "teachable agents" driven by large language models (LLMS) in music theory learning. The experiments showed that this approach could significantly improve academic performance, reduce cognitive burden, and thereby enhance learning efficiency. On the other hand, the review by Merchán Sánchez-Jara et al. [1] also clearly pointed out that the realization of real-time interaction and personalized feedback by AI in music education is a key path to improve learning efficiency and teaching interactivity.

In conclusion, introducing intelligent speech recognition technology into vocal music teaching, especially for enhancing classroom interactivity and learning efficiency, holds both theoretical innovation and practical significance. Moreover, AI technology can promote interaction between teachers and students and enhance learning quality through intelligent feedback. At the same time, it also reminds us to pay attention to maintaining the subjectivity of teaching and the risks of relying on technology. Based on this, this study intends to explore the specific impact of intelligent speech recognition on interactivity and efficiency in vocal music teaching through experimental design, aiming to make up for the empirical deficiency of existing research in the field of vocal music and promote the in-depth application and theoretical construction of intelligent educational technology in art teaching.

2. Literature Review and Theoretical Framework

2.1 Literature Review: Progress in the Integration of Intelligent Speech Recognition and Vocal Music Education

Driven by deep learning, intelligent speech recognition has undergone a paradigm leap from hidden Markov models to deep neural networks/end-to-end acoustic modeling, significantly enhancing recognition accuracy and real-time performance, creating conditions for classroom-level and personalized learning support [9]. In educational scenarios, the value of intelligent speech recognition mainly lies in three types of functions: First, real-time transcription and speech understanding, reducing the burden on teachers' note-taking and students' notes; Secondly, based on the automatic assessment and feedback linkage of pitch/rhythm/articulation, the recognition results are combined with music signal processing (such as fundamental frequency detection, beat alignment) to provide learners with actionable error correction suggestions. Thirdly, data-driven process assessment uses time series data to depict practice frequency, error types and improvement trajectories, thereby supporting teachers' formative evaluation. Educational research shows that timely and specific feedback can significantly enhance learning outcomes and interaction quality, and the linearity and traceability of intelligent speech recognition are precisely in line with this [10].

At the level of learning efficiency, the cognitive load theory points out that the reduction of external load and the optimization of essential load/related load are conducive to improving learning efficiency [11]. Presenting key points of pronunciation, pitch deviations and rhythm errors in structured prompts can reduce the additional load on learners in the process of "error finding - location - understanding", and increase the effective practice density per unit time. Meanwhile, the multimedia learning theory emphasizes that the coordinated presentation of text, speech and visualization can promote processing and retention [12]; Intelligent voice recognition forms a closed loop of "what is sung

- what is seen - what is corrected", which can create a high-frequency and low-cost deliberate practice cycle in vocal training. Regarding interactivity, social constructivism emphasizes achieving capacity leaps within recent development zones through the guidance of more experienced others [13]; The real-time conversational feedback and peer review evidence facilitated by intelligent voice recognition help expand the breadth and depth of classroom interaction. Furthermore, the Technology Acceptance Model suggests that perceived usefulness, ease of use, and social impact will influence teachers' and students' willingness to adopt and their continued use [14]; Therefore, the teaching effect of speech recognition in vocal music classes not only depends on the performance of the algorithm, but is also restricted by usability design and classroom organization strategies. Looking at the existing work, there is a considerable amount of evidence for the application of intelligent speech recognition in language learning and general music education. However, systematic empirical evidence for real-time error correction and process assessment directly oriented to vocal performance is still insufficient. This constitutes the entry point and supplementary value of this study.

2.2 Theoretical Framework and Research Hypotheses

Based on the above review, this study constructs a mechanism model of "intelligent speech recognition empowerment - interaction enhancement - load optimization - efficiency improvement". Its core logic is:

Technology empowerment layer (intelligent speech recognition performance and function): High-accuracy and low-latency recognition and coupled pitch/rhythm analysis, providing immediate, actionable and traceable feedback evidence. This type of evidence is presented in combination with natural language prompts through visualization (error highlighting, curve alignment, suggested segments) (corresponding to the principles of multimedia learning), reducing the search and reasoning costs for learners.

Interactive process layer (Quality of interaction between teachers and students/among students) : Teachers can provide targeted explanations and stratified guidance based on data, while students can engage in dialogue and collaboration around the turn-based task of "evidence - correction - singing again", thereby enhancing the frequency of speaking, the quality of questions, and the speed of feedback response.

Cognitive mechanism layer (load redistribution and motivation stimulation): External loads decrease due to automatic recording and precise positioning; Cognitive processing directly related to the task objective (such as controlling breathing and adjusting resonance) is strengthened [4]. At the same time, clarifying the progress trajectory and enhancing self-efficacy can help strengthen intrinsic motivation and continuous commitment.

Outcome Layer (Learning Efficiency and Learning Achievement): The effective practice duration per unit time, the speed of error correction, the accuracy of singing and the stability of rhythm are improved, forming quantifiable efficiency and quality gains.

Under this framework, research hypotheses are proposed:

• H1 (Interactivity): Compared with traditional teaching, vocal music classes supported by intelligent voice recognition will significantly increase the frequency and quality of classroom interaction (such as the timeliness, pertinence and operability of feedback).

• H2 (Learning Efficiency): The intervention of intelligent speech recognition can significantly enhance learning efficiency (the proportion of correct singing within a unit of time, error location and correction speed).

• H3 (Mediating Mechanism): The improvement of interaction quality and the optimization of cognitive load play a positive mediating role between intelligent speech recognition and the enhancement of learning efficiency.

• H4 (Moderating Effect): The learner's basic level, the difficulty of the repertoire, and the system availability (perceived ease of use/usefulness) moderate the teaching effect of intelligent speech recognition. The gains are more obvious in high availability and medium-difficulty scenarios.

Based on this, subsequent research will conduct measurements and path analysis around the index chain of "interactivity - load - efficiency - achievement" in the experimental group (intelligent speech recognition + vocal music teaching) and the control group (traditional teaching), and verify the mediating and moderating paths in combination with process data. This framework not only connects with the

classic theories of educational psychology and learning science, but also incorporates the engineering feasibility of speech recognition into the teaching design, ensuring that the results have both theoretical explanatory power and practical applicability.

3. Research Methods and Experimental Design

3.1 Research Subjects and Sample Selection

The experimental subjects of this study were 60 students majoring in vocal music from the first to the third year of undergraduate studies at the music college of a comprehensive university. After stratified sampling by gender, grade and singing level, they were randomly assigned to the experimental group and the control group, with 30 students in each group. Choosing to study vocal music at the undergraduate level has two advantages: First, they have received systematic vocal music training and possess a certain professional foundation, which makes it easier for them to observe the improvement effect under technical intervention. Secondly, this group is currently at a crucial stage in cultivating their singing skills and stage presence, and the changes in interaction quality and learning efficiency are more representative. To avoid sample bias, this study excluded students whose training effects were affected by factors such as voice disorders and hearing impairments. Meanwhile, to control for the differences in teacher factors, in this study, the same vocal music teacher with the title of associate professor was responsible for teaching both groups of courses. This teacher has more than 10 years of vocal music teaching experience, and has a certain exploration foundation in information-based teaching, and can skillfully operate an intelligent speech recognition system.

3.2 Teaching Experiment Design

This study adopted a quasi-experimental design. The experimental period was 12 weeks, with two classes per week, each lasting 90 minutes.

Control Group: The traditional vocal music teaching model was adopted. Teachers provided vocal and singing guidance through piano accompaniment and oral explanations, and students made corrections based on teacher feedback and personal perception.

Experimental Group: On the basis of traditional teaching, an intelligent speech recognition system was introduced. This system integrates acoustic signal processing (pitch detection, rhythm comparison, timbre analysis) and speech recognition modules, capable of real-time transcription and pitch and rhythm analysis of students' singing, and providing immediate visual feedback. Teachers use the visual reports output by the system for targeted explanations, while students can conduct self-diagnosis and repetitive practice through system playback and marked information.

To ensure the consistency of teaching content, both groups adopted the same vocal music textbooks and pieces, covering both Chinese art songs and classic Western art songs, to ensure a similar distribution of difficulty.

3.3 Data Collection Methods

To comprehensively examine the role of intelligent speech recognition in vocal music teaching, this study adopts a diversified data collection method:

3.3.1 Classroom observation and interaction records

The number of interactions between teachers and students in the classroom, the frequency of students' questions and responses, as well as the timeliness and specificity of teachers' feedback were recorded through video and observation scales.

3.3.2 Learning efficiency test

Standardized vocal tests were conducted before and after the experiment, covering four dimensions: pitch accuracy, rhythm, articulation and emotional expression, and were independently scored by three vocal experts with the title of associate professor or above. At the same time, the students' singing recordings were collected and the intelligent voice recognition system for objective analysis was used to obtain quantitative indicators such as average pitch deviation and rhythm error rate.

3.3.3 Questionnaire survey

The "Questionnaire on Interaction and Learning Experience in Vocal Music Classroom" was designed, covering dimensions such as interaction perception, learning efficiency perception, and technical acceptance. The questionnaire adopted the Likert five-point scale to ensure reliability and validity. Internal consistency was tested by Cronbach's α test.

3.3.4 Interview

Semi-structured interviews were conducted with 10 students and teachers to collect their

subjective feelings and suggestions regarding the application of intelligent speech recognition.

3.4 Research Variables and Measurement Indicators

To verify the theoretical framework and research hypotheses, this study mainly measures the following variables:

Independent variable: Teaching mode (Traditional vs. intelligent speech recognition assistance).

Dependent variable:

Learning interactivity: The number of classroom speeches, the frequency of teacher-student dialogues, the timeliness and pertinence of feedback.

Learning efficiency: The correct proportion of singing per unit time, the extent of improvement in pitch and rhythm accuracy, and the speed of task completion.

Mediating variable: Cognitive load (measured by the subjective scale NASA-TLX and behavioral indicators).

Moderating variables: Learner's basic level, difficulty of the repertoire, and system availability awareness.

Table 1 shows the main variables and measurement methods.

Table 1. Types of Research Variables and Measurement Indicators

Variable type	Indicator	Measuring tools/methods
Learning interactivity	Number of speeches, quality of feedback, and duration of interaction	Classroom video + encoded scale
Learning efficiency	Pitch deviation, rhythm error rate, and singing score	Expert scores + quantified data from the speech recognition system
Cognitive load	Subjective load, time pressure, and frustration	NASA-TLX Scale
Technical acceptance	Perceived usefulness, perceived ease of use, willingness	Questionnaire items based on TAM/UTAUT

4. Empirical Results and Analysis

4.1 Descriptive Statistics and Reliability and Validity Tests

A total of 60 questionnaires were distributed in this study, and 60 valid questionnaires were retrieved, with an effective recovery rate of 100%. Among them, there were 30 people in the experimental group and 30 in the control group. The gender ratio was close (40% male and 60% female), and the average age was 19.8 years old.

In terms of the reliability of the questionnaire, the overall Cronbach's α coefficient of the "Vocal Music Classroom Interaction and Learning Experience Questionnaire" was 0.913, and all subscales were greater than 0.85, indicating a relatively high internal consistency. The validity test results showed that the KMO value was 0.871, and the Bartlett sphericity test was significant ($p < 0.001$), indicating that the scale had good structural validity and was suitable for factor analysis.

4.2 Experimental Results of Classroom Interactivity

Through classroom video coding and questionnaire surveys, data on the frequency and quality of interaction between teachers and students were obtained.

Table 2. Comparison Results of Classroom Interactivity (M ± SD)

Indicators	Experimental Group (n=30)	Control group (n=30)	t value	p value
The average number of speeches made by students	7.83 ± 2.11	4.25 ± 1.87	6.54	<0.001
The number of interaction rounds between teachers and students	12.67 ± 3.05	7.92 ± 2.34	6.02	<0.001
Teacher feedback timeliness scoring	4.56 ± 0.48	3.72 ± 0.61	5.98	<0.001
Feedback targeted scoring	4.71 ± 0.42	3.68 ± 0.55	7.01	<0.001

Table 2 shows that the average number of times students in the experimental group spoke was significantly higher than that in the control group, indicating that with the assistance of intelligent speech recognition, students were more willing to express themselves and try singing. The difference in the number of interaction rounds between teachers and students is also quite significant. The interaction frequency of the experimental group is nearly 1.6 times that of the

control group, which indicates that the real-time feedback mechanism of intelligent speech recognition effectively promotes the bidirectionality of classroom communication. In addition, in terms of the quality of teachers' feedback, the experimental group was significantly superior to the control group in both timeliness and pertinence. This result means that teachers can quickly identify students' problems in pitch and rhythm through the system and provide more targeted feedback, thereby enhancing the teaching efficiency and interaction depth in the classroom. Overall, the results in Table 2 reveal that intelligent speech recognition not only enhances students' classroom participation but also optimizes the interaction mode between teachers and students, making teaching activities more efficient, dynamic and precise.

4.3 Experimental Results on Learning Efficiency

The learning efficiency is evaluated through the scores of vocal music experts and the quantitative data from the intelligent speech recognition system.

Table 3. Comparison Results of Learning Efficiency (M ± SD)

Indicators	Experimental Group (n=30)	Control group (n=30)	t value	p value
Average pitch deviation (cent)	18.6 ± 4.72	32.5 ± 6.91	-8.64	<0.001
Rhythm error rate (%)	6.7 ± 2.35	12.3 ± 3.18	-7.56	<0.001
Expert comprehensive score (out of 100)	87.4 ± 5.82	78.9 ± 6.77	5.43	<0.001
Single task completion time (seconds)	132.8 ± 18.5	158.6 ± 21.3	-4.67	<0.001

Table 3 shows that the experimental group is significantly superior to the control group in all indicators of learning efficiency. Firstly, in terms of the average pitch deviation, the error of the experimental group was reduced by approximately 43% compared to the control group, indicating that the immediate feedback provided by intelligent speech recognition helps students quickly identify and correct pitch problems. Secondly, the rhythm error rate in the experimental group dropped to 6.7%, significantly lower than 12.3% in the control group, indicating that the system also has a significant auxiliary effect on rhythm training. In terms of overall performance, the expert scores of the students in the experimental group were approximately 8.5 points higher than those in the control group, further verifying their advantage in overall singing quality. Finally, in terms of the time required to complete the single task, the experimental group took less, indicating a substantial improvement in learning efficiency. Overall, the data in Table 3 verify the promoting effect of intelligent speech recognition on learning efficiency.

4.4 Analysis of the Mediating Effect of Cognitive Load

The mediating effect was tested using the PROCESS macro, and the results showed that:

The direct effect of the teaching mode on learning efficiency is significant ($\beta = 0.48$, $p < 0.001$); The negative effect of the teaching mode on cognitive load was significant ($\beta = -0.36$, $p < 0.01$), that is, the subjective load of the students in the experimental group was lower. The negative prediction of cognitive load on learning efficiency was significant ($\beta = -0.42$, $p < 0.01$); The Bootstrap test results showed that the indirect effect was significant (95% CI: 0.08-0.23), indicating that cognitive load played a partial mediating role between intelligent speech recognition and learning efficiency.

4.5 Analysis of the Moderating Effects of Technology Acceptance

Taking learners' perceived ease of use and usefulness as moderating variables, the influence on the teaching effect of the experimental group was examined. The results show that

When students' acceptance of technology is high, the learning efficiency improvement effect of intelligent speech recognition is significantly enhanced ($\beta = 0.29$, $p < 0.05$);

When students' acceptance is low, although the effect of intelligent speech recognition is still positive, the gain is significantly weakened.

This indicates that the acceptance of technology plays a moderating role in the teaching effectiveness of intelligent speech recognition applications, verifying the theoretical expectations of the TAM and UTAUT models.

4.6 Qualitative Analysis of Interview Results

Through thematic analysis of semi-structured interviews with 10 students and teachers, the

following viewpoints are summarized:

Positive experience: Students generally believe that "real-time feedback", "clear error location" and "convenient repetitive practice" are the greatest advantages. Teachers believe that the system "improves classroom efficiency" and helps them "understand students' problems more accurately".

Challenges and shortcomings: Some students indicated that the system's recognition accuracy declined in noisy environments. Teachers are concerned that technological reliance might weaken students' ability to listen and distinguish independently.

Improvement suggestions: It is hoped that the system can add functions for emotional expression and timbre analysis, and that the interface design be more in line with the music teaching context.

4.7 Summary

Based on the comprehensive quantitative and qualitative results, this study finds that intelligent speech recognition technology can significantly enhance the interactivity and learning efficiency of vocal music classes. Cognitive load plays a partial mediating role between the two, and technical acceptance plays a moderating role therein. Students and teachers generally hold a positive attitude, but in practical application, attention should be paid to the issues of identification accuracy and teaching balance.

These results not only verify the research hypotheses proposed earlier, but also provide empirical support for the in-depth application of intelligent speech recognition in the field of music education.

5. Conclusion

This study takes the application of intelligent speech recognition in vocal music teaching as the entry point. Through quasi-experimental design and multi-dimensional data collection, it systematically examines the impact of this technology on classroom interactivity and learning efficiency. The research results show that intelligent speech recognition can significantly enhance students' participation and interaction frequency in the classroom, and improve the timeliness and pertinence of teachers' feedback. In terms of learning efficiency, the students in the experimental group were significantly superior to those in the control group in terms of pitch accuracy, rhythm, singing quality and task completion speed. Further analysis shows that cognitive load plays a partial mediating role between teaching mode and learning efficiency, while technology acceptance has an important moderating effect on the exertion of the effect.

From a practical perspective, the introduction of intelligent speech recognition not only provides immediate feedback and visual analysis tools for vocal music teaching, but also promotes the improvement of classroom interactivity and learning autonomy. However, the technology still faces challenges such as recognition accuracy and learner dependence in practical applications, and it is necessary to further optimize the system functions and combine them with the professional judgment of teachers. Future research can verify its applicability in larger samples, diverse repertoire and interdisciplinary scenarios, and explore higher-level intelligent support such as emotional expression and artistic expression to achieve innovation in music education under the integration of human and machine.

References

[1] Merchán Sánchez-Jara J F, González Gutiérrez S, Cruz Rodríguez J, Syroyid Syroyid B. Artificial Intelligence-Assisted Music Education: A Critical Synthesis of Challenges and Opportunities. Education Sciences, 2024, 14(11): 1171.

[2] Zhang T Y. Research on Online Vocal Music Smart Classroom-Assisted Teaching Based on Wireless Network Combined With Artificial Intelligence. International Journal of Web-Based Learning and Teaching Technologies, 2024, 19(1).

[3] Sun Q. Large Model-Driven Intelligent Learning Systems: Development Trends and Key Technologies. Research on artificial intelligence education, 2025, 1(1): 27-40.

[4] Wang S S. The Application of AI Intelligent Speech Recognition Technology in Primary School English Listening and Speaking Teaching. Chinese teacher. 2024, 22.

[5] Ma Y L. The Current Status and Future Prospects of the Application of Artificial Intelligence in the Field of Children's Education. Environment and Development, 6(12): 312-314.

[6] Zhou Y. A Comparative Study of the Impact of Artificial Intelligence on the Ecological Niches of English and Japanese Majors from

the Perspective of Educational Ecology. Advances in Education, 2025, 15(8): 1069-1078.
[7] Li P P, Wang B. Artificial Intelligence in Music Education. International Journal of Human-Computer Interaction, 2024, 40(16): 4183-4192.
[8] Jin L X, Lin B C, Hong M Z, Zhang K, So H J. Exploring the Impact of an LLM-Powered Teachable Agent on Learning Gains and Cognitive Load in Music Education. 2025, arXiv:2504.00636.
[9] Li W C. Research on Music Education Simulation Based on Interactive Experience of Virtual and Reality. Systems and Soft Computing, 2025, 7: 200343.
[10] Cheng L. The Impact of Generative AI on School Music Education: Challenges and Recommendations. Arts Education Policy Review, 2025: 1-8.
[11] Després J P, Dubé F. The Music Learner Voice: A Systematic Literature Review and Framework. Frontiers in Education, 2020, 5: 119.
[12] Wu D, Li H, Chen X. Analysis of the Impact of General Large Models of Artificial Intelligence on Education Applications. Open Education Research, 2023, 29(2): 19-25.
[13] Dai Y. Generative Artificial Intelligence Technology Empowers teachers' classroom Teaching Quality Assessment, Advances in Education, 2024, 14(6): 1264-1271.
[14] Rexhepi F G, Breznica R K, Rexhepi B R. Evaluating the Effectiveness of Using Digital Technologies in Music Education. Journal of Educational Technology Development and Exchange, 2024, 17(1): 273-289.

