Eng Int Syst (2025) 3: 249–261

Engineering
Intelligent Systems

Design and Implementation of an Automatic Evaluation System for English-Chinese Interpretation Based on Artificial Intelligence

Fang Ren*

Xi'an Fanyi University, Xi'an, 710105 China

With the development of globalization, English-Chinese interpretation plays an increasingly important role in international communication. The aim of this study is to design and implement an artificial intelligence-based automatic evaluation system for English-Chinese interpreting, and to provide an objective and efficient tool for the evaluation of interpretation quality. By using a hybrid model comprising a constitutional neural network (CNN) and a long short-term memory network (LSTM), the system can effectively deal with complex linguistic phenomena in interpreting content. In this study, through comprehensive data collection and re-processing, high-quality data was obtained as input for model training. System tests show that the model achieves good results in terms of functionality, performance and user experience, especially in regard to semantic accuracy and fluency. However, the study also has limitations, such as the diversity of data samples and the processing power in complex contexts. Future research will focus on expanding the datasets and further optimizing the algorithm to improve the generalization ability and accuracy of the system. In general, this study provides a new technical perspective for the field of automatic assessment of English-Chinese interpreting, and provides a valuable reference for related research.

Keywords: artificial intelligence, English-Chinese interpretation, automatic evaluation system, convolution neural network, long short-term memory network

1. INTRODUCTION

In the context of globalization, cross-cultural communication is becoming increasingly frequent. As an important means of communication, English-Chinese interpretation plays a key role in international business, diplomacy, education and other fields. With the rapid development of artificial intelligence (AI) technology, especially the progress of natural language processing (NLP) and machine learning, the automatic evaluation system of English-Chinese interpretation has become a research and application hotspot. This system aims to evaluate the quality of interpretation objectively and efficiently, which is expected to solve the problems of subjectivity and poor efficiency of traditional evaluation methods. Although AI has made remarkable progress in speech recognition and language understanding, it still faces many challenges in the field of automatic evaluation of English-Chinese interpretation. For example, it needs to accurately understand and evaluate the semantic accuracy, fluency, and cultural adaptability of an interpretation, and also needs to deal with non-standard words and dialects in spoken language. Additionally, the diversity and subjectivity of interpretation assessment pose challenges for the development of uniform assessment criteria.

In the study of the automatic evaluation system of English and Chinese interpretation, many scholars have discussed the standards, methods and tools of interpretation evaluation.

*Corresponding author's e-mail: renfang168@outlook.com

Su [1] emphasized the importance of assessment standards in enhancing students' awareness of interpreting skills, which has direct guiding significance for the accuracy and reliability of automatic assessment systems. At the same time, Su [2] explored the differences in the assessment of interpretation by teachers with different backgrounds, highlighting the subjectivity and diversity of assessment criteria. These studies reveal the limitations of traditional interpretation evaluation methods and provide a background for the development of automatic evaluation systems. In addition, Irribarra and Arneson [3] challenged the definition and interpretation of dimensions in educational and psychological assessment, which has implications for how assessment criteria can be quantified and defined in interpretation assessment systems. Han [4] studied the influence of linguistic background and orientation on interpretation evaluation, and further emphasized the importance of considering linguistic and cultural factors when developing interpretation evaluation systems. In terms of technology, Lu and Han [5] explored the automatic evaluation of spoken translation based on the evaluation indicators of machine translation, providing a reference for this study on how to use existing technologies to evaluate interpretation. In addition, Han and Zhao [6] examined the accuracy of peer review in interpreting quality assessment, also providing a valuable comparison for the automatic assessment method adopted in this study. The existing literature provides a rich theoretical and practical basis for this study. On the basis of these studies, this study aims to improve the quality of interpretation teaching and practice by utilizing advanced AI technology to solve the subjectivity and efficiency problems in traditional interpretation assessment.

The main goal of this study is to create an artificial intelligence-based automatic evaluation system for English-Chinese interpretation. This system aims to provide an objective, efficient, and accurate interpretation evaluation tool using advanced machine learning algorithms and natural language processing techniques. Additionally, the research aims to explore the potential of AI in language conversion and cultural adaptation assessment to reduce the subjectivity and inconsistencies present in traditional interpretation evaluation.

The research has several significant aspects. Firstly, it introduces a new methodology for evaluating interpretation quality. Traditional interpretation evaluation is often based on personal experience and subjective judgment; however, the system proposed in this study can provide a more objective and consistent evaluation standard. Secondly, this system can significantly improve the efficiency of assessment, especially in scenarios that require a lot of assessment, such as interpretation training and teaching. Thirdly, by analyzing a large number of interpreting samples, the system can also identify common problems and trends in interpreting, providing valuable feedback for interpreting teaching. This study not only introduces a new technical means for the field of interpretation evaluation, but also provides an empirical study on the potential of AI in a wider range of language processing applications. This is helpful for promoting the application of AI technology in education, communication, and other fields.

This research focuses on the design and implementation of an automatic evaluation system for English-Chinese interpretation based on artificial intelligence. At the heart of the system is the use of the latest machine learning and natural language processing technology to automatically assess the quality of interpretation. The research will involve several key steps: The research will conduct in-depth data collection and re-processing work. This includes collecting English-Chinese interpretation samples from multiple sources and performing the necessary cleaning and formatting of these samples for subsequent machine learning training and evaluation. The research will focus on model construction and algorithm experiments. This involves selecting the right machine learning algorithm, training and optimizing the algorithm to ensure the accuracy and efficiency of the evaluation. The process also includes comparing and analyzing the performance of different algorithms and models so as to select the best solution. The research will also focus on the implementation of the system.

The research involves integrating algorithms, designing system architecture, and creating a user-friendly interface to make it easy for non-expert users to evaluate interpretations. Testing and evaluation are crucial parts of the research, and include the comprehensive testing of the system's function and performance to ensure its stability and reliability. Additionally, user feedback will be collected to assess the practicality of the system and further optimize it. In summary, the aim of this study is to provide an innovative solution for evaluating English-Chinese interpretation using advanced artificial intelligence technology, while also serving as a valuable reference for related research fields.

2. THEORETICAL BASIS AND TECHNICAL BACKGROUND

2.1 Artificial Intelligence and Natural Language Processing

In this study, the theoretical foundation and technical background of artificial intelligence (AI) and natural language processing (NLP) are the core parts. Artificial intelligence, especially its application in natural language processing, provides powerful technical support for understanding and generating human language [7]. The rapid development of this field enables machines to process and understand complex linguistic phenomena more effectively, which is crucial for the construction of an automatic evaluation system for English-Chinese interpreting.

Natural language processing (NLP) is a technology that uses the computer as a tool to process natural language information in written and spoken form. Natural language processing technology enables computers to parse, understand and generate human language, including grammar analysis, semantic understanding, emotion analysis and other language characteristics [8, 9]. In the interpretation evaluation system, NLP technology is used to evaluate the semantic correctness and fluency of the interpretation content. For example, semantic analysis is used to determine whether the translation is accurate, and grammatical analysis is used to evaluate whether the structure of the sentence conforms to the language norms of the target language.

The use of machine learning, especially deep learning, in the field of natural language processing (NLP) is a key aspect of this research. Deep learning techniques such asconvolution neural networks (CNN) and recurrent neural networks (RNN) are widely applied in tasks such as speech recognition, text classification, and translation quality assessment [10]. These technologies can significantly enhance the performance of the interpretation evaluation system, making it more accurate and efficient in handling complex linguistic phenomena.

In summary, the theories and techniques of artificial intelligence and natural language processing not only form the technical foundation of this study [11], but also offer the potential for creating an efficient and accurate automatic evaluation system for English-Chinese interpretation.

2.2 Standards and Methods of Interpretation Evaluation

It is very important to understand and integrate the existing evaluation standards and methods in constructing an artificial intelligence-based automatic evaluation system for English and Chinese interpretation. The evaluation of interpretation involves several criteria: semantic accuracy, fluency of expression, accurate use of language and cultural adaptability. Together, these criteria constitute a comprehensive framework for assessing the quality of interpretation [12].

Semantic accuracy is the core of evaluating the quality of interpretation, which requires that the content of interpretation must be faithful to the original text to ensure the accurate transmission of information [13]. Fluency refers to the coherence and naturalness of interpretation and indicates whether the interpreter can express the meaning of the original text fluently and clearly. The accurate use of language involves the correctness of vocabulary, grammar and pronunciation, and emphasizes the standardization of interpretation content in linguistic form. Cultural adaptability refers to the ability of interpreters to convey cross-cultural information, and examines whether interpreters can deal with language expressions appropriately in different cultural environments.

The existing interpretation evaluation methods usually include expert review, peer evaluation and self-evaluation. Expert review relies on the experience and judgment of professional judges and attaches importance to the professionalism and comprehensiveness of the assessment. Peer review focuses more on the mutual evaluation of interpreters to promote learning and improvement. Self-assessment encourages interpreters to reflect upon and evaluate their own performance [14, 15]. However, these methods are often prone to subjectivity and are less efficient in large-scale assessments.

The goal of this study is to integrate traditional evaluation criteria and methods with modern artificial intelligence techniques. This integration will lead to the development of a system that can automatically, objectively, and efficiently assess the quality of interpretation. This approach aims to improve the accuracy and efficiency of assessment and provide valuable feedback and guidance for the teaching and practice of interpretation tasks.

2.3 Challenges of Technology Convergence

The development of an AI-based automatic evaluation system for English-Chinese interpretation faces several challenges in regard to technology integration. These challenges stem primarily from the complexity of applying advanced artificial intelligence techniques, particularly natural language processing (NLP) and machine learning (ML), to the specific field of interpretation evaluation [16].

The first challenge is the diversity and complexity of languages. The interpretation of content usually contains rich context, diverse expressions and culture-specific elements, which makes it significantly more difficult for machines to understand and evaluate. For example, the same sentence may have different meanings in different cultures and contexts, which requires the evaluation system to have a high context recognition ability [17, 18].

The accuracy and adaptability of the algorithm is also an important challenge. Assessing interpretation quality is not only about identifying linguistic structures, but also about understanding the intentions and emotions expressed by the translator. Therefore, selecting and optimizing algorithms suitable for interpretation evaluation is the key to achieving efficient and accurate evaluation [19]. This requires substantial data support and in-depth algorithm adjustment in order to adapt to the characteristics of interpretation.

In addition, the usability and user friendliness of the system are also important considerations. An ideal interpretation evaluation system should be technologically advanced as well as convenient and easy to use in practical applications [20]. This requires that the user's interactive experience be fully considered in the system design to ensure that the system's interface is intuitive and easy to navigate.

The quality of data and ethical issues are also important aspects that must be addressed when technology convergence occurs. High-quality data is a prerequisite for training effective algorithms, and privacy protection and ethical norms must be strictly observed during data collection and processing [21].

To sum up, while technology convergence offers innovation and improvements in efficiency, it also brings a series of challenges. Only by comprehensively considering these challenges and adopting corresponding solutions can we effectively realize the application of artificial intelligence technology in the field of interpretation evaluation.

3. DATA COLLECTION AND TELEPROCESSING

3.1 Data Source and Acquisition Policy

In this study, data collection and re-processing provide the basis for building an efficient automatic evaluation system for English-Chinese interpreting. To ensure the quality and diversity of data, the data sources and acquisition policies are established as outlined in Table 1.

In this study, the data acquisition strategy comprises the following steps:

Table 1 Data Sources and Acquisition Policies.

Data source class	Specific source	Description	Data type	Purpose
Open database	Recording of the International Interpretation Conference	Audio recordings and transcripts of English-Chinese interpretation from international conferences	Audio and text	Language recognition and translation quality assessment for training and validating models
Educational institution	Sample mock interpretation provided by the Language Learning Centre	Audio recordings and transcripts of simulated interpreting exercises conducted by students and teachers	Audio and text	Used to analyze the development of interpreting skills and common types of errors
Online platform	Web translation forums and communities	Sample interpretation shared by professional and amateur translators	text	It is used to increase the diversity of samples and practical application scenarios
Cooperative agency	Actual interpretation materials provided by partner companies and organizations	Including interpretation materials for practical scenarios such as business meetings and seminars	Text and audio	It is used to test the accuracy and reliability of the system in practical applications

Table 2 Data Preprocessing Process.

Procedure	Description	Purpose	Tools/Methods
Data cleaning	Remove invalid, incomplete, or irrelevant data records	Ensure data quality and improve the accuracy of model training	Automated script, manual audit
Format standardization	Convert all data to a unified format (e.g., unified audio quality, text encoding)	Facilitate data processing and analysis	Audio editing software, text processing tools
Text transfer	Convert audio data to text	For text analysis and model training	Automatic speech recognition (ASR) system
Data annotation	Annotate textual data (e.g. semantic labels, syntactic structures)	For training more accurate NLP models	Manual annotation, collaboration tools
Data enhancement	Increase data diversity through technical means (e.g. synthetic accent variations)	Enhance the generalization ability of the model	Data enhancement tool
Data partitioning	The data set is divided into a training set, a validation set, and a test set	For model training and evaluation	Data partition script

(1) Cooperation Agreements and ethical reviews: Establish partnerships with educational institutions and partner institutions and ensure that all data collection complies with ethical reviews and privacy protection standards.

(2) Quality screening and classification: Evaluate the quality of the collected data to ensure that the audio is clear and the text is accurate, and classify according to the purpose.

(3) Format unification and teleprocessing: Convert all data into a unified format, such as audio into the corresponding text format, to ensure the consistency of data teleprocessing.

By using diversified data sources and applying rigorous acquisition strategies, high-quality and representative data can be collected for this study, providing a solid basis for subsequent model training and system evaluation.

3.2 Data Teleprocessing Process

In this study, data teleprocessing is a key step to ensure the accuracy and effectiveness of an automatic evaluation system for English-Chinese interpreting. The data teleprocessing process is outlined in Table 2.

When teleprocessing the data, the main focus is on ensuring the accuracy, consistency, and representation of the data. Fundamental steps such as data cleaning and formatting standardization are crucial to ensure the initial quality of the data. Additionally, tasks like text translation and data annotation are important for adapting the data for NLP models and to provide rich information for training and analysis. The aim of data enhancement is to improve the model's robustness in handling different accents and expressions. Lastly, reasonable data partitioning is key to ensuring the effectiveness of model training and the accuracy of evaluation.

By carefully carrying out these teleprocessing steps, we can ensure that the collected data effectively supports subsequent model training and system evaluation. This establishes a solid foundation for the development of an efficient and accurate interpretation evaluation system.

3.3 Security and Ethics of Data

Ensuring the security and ethics of data is a crucial part of this study, especially given that interpretation data may contain sensitive or personal information. The strategies for data security and ethics are outlined in Table 3.

When implementing these strategies, the focus is on ensuring that all data processing activities comply with the highest legal and ethical standards. Compliance audits ensure that the study complies with all relevant laws and regulations. Data is anonymized and stored encrypted to protect the security and privacy of personal information. Access control and security audits are designed to prevent unauthorized access and disclosure of data. Finally, the research team is trained in ethics to increase their sensitivity and responsibility for data processing.

These comprehensive measures ensure ethical compliance and the safety of the data, while also enhancing the overall quality of the study and strengthening its credibility.

4. ALGORITHM DESIGN AND EXPERIMENT OPTIMIZATION

4.1 Algorithm Selection and Theoretical Support

Given the focus of this study, the choice of algorithm is very important. This section discusses the selection of suitable machine learning algorithms for handling the complex task of evaluating interpretations. The following is an analysis of several potential algorithms and their theoretical support,

Table 3 Data Security and Ethical Strategies.

Strategy	Description	Implementation method	Purpose
Compliance review	Ensure that data collection and processing comply with relevant laws and regulations	Legal advice, ethical review	Comply with data protection laws such as GDPR
Data randomization	Remove or replace personally identifiable information	Data desensitization tools, manual processing	Protect personal privacy and reduce the risk of disclosure
Encrypted storage	Encrypt sensitive data	Encryption technology and secure storage solutions	Prevents unauthorized access to data during storage and transmission
Access control	Restrict access to data	Access control system, permission management	Ensure that only authorized personnel have access to the data
Security audit	Periodically review and evaluate data security	Security audit tools, professional audit services	Find and correct security vulnerabilities to improve system security
Ethical training	Train research teams in data ethics and protection	Training seminars, online courses	Increase team awareness of data protection and ethics

using mathematical modeling to describe the core principles of each algorithm.

(1) Conventional Neural Network (CNN):

Core principle: CNNS are suitable for processing data with a grid structure, such as time series or image data. When evaluating interpretation, CNNS can be used to extract the features of speech signals.

The mathematical model is shown in (1):

$$f(x) = ReLU(W * x + b) \tag{1}$$

where x is the input data, W and b are the weight and bias of the convolution layer, respectively, $ReLU$ is the activation function, and $*$ represents the convolution operation.

(2) Recurrent neural Network (RNN):

Core principle: RNNS are particularly suited for the processing of sequential data and are able to capture time series information when processing language data.

The mathematical model is shown in (2):

$$h_t = tanh(W_{hh}h_{t-1} + W_{xh}x_t + b_h) \tag{2}$$

where h_t is the hidden state of time step t, h_{t-1} is the hidden state of the previous time step, x_t is the input of time step t, and W_{hh}, W_{xh} and b_h are network parameters.

(3) Long Short-term Memory Network (LSTM):

Core principle: LSTM is a variant of RNN that addresses the gradient disappearance problem of traditional RNN in long sequence data processing. It is well-suited for processing long sentences in sentiment analysis.

The mathematical model is shown in (3):

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f) \tag{3}$$

where f_t is the output of the forgetting gate, σ is the sigma activation function, W_f and b_f are the weight and bias of the forgetting gate, h_{t-1} is the hidden state of the previous time step, and x_t is the input of the current time step.

When choosing the most suitable algorithm, taking into account the characteristics of the interpreting content, such as the complexity and diversity of the language, LSTM may be the most suitable choice due to its ability to deal with long-term dependencies. In addition, a hybrid model combining CNN and LSTM, where CNN is used to extract speech features and LSTM processes sequence data, may be more effective, together enabling more accurate evaluation of the interpretation.

The comprehensive application of this algorithm allows complex interpretation data to be dealt with more effectively, thereby improving the accuracy and reliability of the evaluation system.

4.2 Experimental Scheme and Implementation Process

Based on the selected algorithms, mainly hybrid models combined with conventional neural networks (CNN) and long short-term memory networks (LSTM), the experimental protocol applied in this study is used to evaluate the effectiveness of these algorithms in the automatic evaluation of English-Chinese interpreting.

(1) Data preparation: The collected and re-processed data are used as the basis for the experiment. The data

are divided into training sets, verification sets, and test sets.

(2) Model design: A hybrid model combining CNN and LSTM, is used to extract and process the features of interpretation data. The model is shown in (4) and (5):

$$CNN_{output} = ReLU(W_{cnn} * x + b_{cnn}) \quad (4)$$

$$LSTM_{output} = LSTM(CNN_{output}) \quad (5)$$

where x is the input data, W_{cnn} and b_{cnn} are the weight and bias of the CNN, * represents the convolution operation, $ReLU$ is the activation function, and $LSTM()$ represents the processing of the LSTM network.

(3) Training and optimization: The training set is used to train the model, and the verification is performed on the verification set to adjust the hyperparathyroidism, such as learning rate, number of layers, etc. Optimization strategies include the use of cross entropy loss functions and stochastic gradient descent (SGD) or Adam optimizers.

The loss function is shown in (6):

$$Loss = -\sum(y\log(\hat{y})) \quad (6)$$

where y is the true label and $\hat{y}$ is the model prediction value.

(4) Evaluation and testing: The performance of the model is evaluated on the test set, and the accuracy rate, recall rate, F1 score and other indicators are used for comprehensive evaluation.

(5) Iterative optimization: According to the test results, the model is interactively optimized, which involves adjusting the model structure and optimizing the algorithm parameters.

(6) Result analysis: Analyze the performance of the model on different types of interpretation data, and identify the strengths and weaknesses of the model.

The study used the experimental scheme mentioned above to thoroughly evaluate the effectiveness of the selected algorithm in the task of automatic interpretation evaluation. The experimental results offer crucial insights into the model's performance, facilitating further optimization of the algorithm and system improvement.

4.3 Performance Analysis and Optimization Strategy

After the algorithm experiment involving the automatic evaluation system for English and Chinese interpretation, it is very important to analyze the performance of the model and formulate the optimization strategy.

(1) Performance evaluation indicators:

The following standard metrics are used to evaluate model performance:

Accuracy is shown in (7):

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (7)$$

The Recall rate is shown in (8):

$$Recall = \frac{TP}{TP + FN} \quad (8)$$

The Precision is shown in (9):

$$Precision = \frac{TP}{TP + FP} \quad (9)$$

F1 scores are shown in (10):

$$F1 = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall} \quad (10)$$

where *TP, TN, FP* and *FN* represent true cases, true negative cases, false positive cases and false negative cases respectively.

(2) Performance analysis method:

A detailed analysis is conducted of the model's performance on the test set, including overall performance and classification performance.

Identify differences in the model's performance on specific types of data (e.g., different accents, different speeds of interpretation, etc.).

(3) Optimization strategy:

Adjust model parameters such as learning rate, batch size, number of hidden layer units, etc.

Algorithmic optimization: improve the model structure (increasing/decreasing the number of layers), using more advanced optimizers (such as Adam).

Data enhancement: Increase the diversity of data samples, such as audio data enhancement by changing speech speed and accent.

(4) Optimize implementation and evaluation:

After implementing the optimization strategy, retrain the model and evaluate the performance improvement.

Use the same performance metrics to compare the difference before and after optimization.

The comparison of model performance before and after optimization is shown in Figure 1.

By means of a comprehensive performance analysis and optimization strategy, this study will be able to significantly improve the accuracy and reliability of the automatic evaluation system for English-Chinese interpretation.

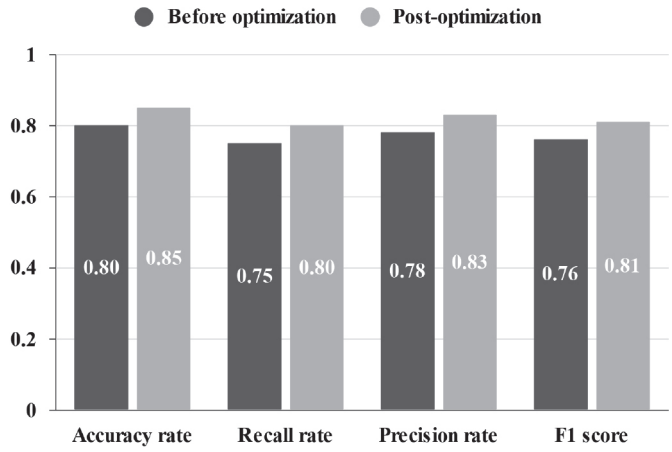

Figure 1 Comparison of model performance.

Table 4 System Components and Functions.

Module	Feature	Description
Data processing layer	Data collection and teleprocessing	Process and prepare data for model training
Model training layer	Algorithm implementation and training	Implement and train a hybrid model
Application layer	User interaction and evaluation output	Provide a user interface to display the evaluation results

5. SYSTEM DEVELOPMENT AND IMPLEMENTATION

5.1 System Architecture and Design Principles

In the development and implementation of an automatic evaluation system for English-Chinese interpretation based on artificial intelligence, the design of the system architecture is very important.

(1) Overview of system architecture

The system adopts a layered architecture comprising a data processing layer, a model training layer and an application layer.

The data processing layer is responsible for collecting and reprocessing data.

The model training layer applies the selected algorithm to train and optimize the model.

The application layer provides the user interface to achieve the interaction and output of interpretation evaluation.

(2) Design principles

Modular: For easy maintenance and upgrade, each part of the system is a separate module.

Scalability: The system design should support future expansion, such as the addition of new algorithms or features.

User friendliness: The design of the application layer should focus on user experience, and be simple and intuitive.

Security and privacy protection: Systems must be designed to ensure data security and user privacy.

(3) System components and functions

As shown in Table 4.

(4) Technical line

Data processing layer: Python, database management system (such as MySQL).

Model training layer: TensorFlow or PyTorch, CUDA (for GPU acceleration).

Application layer: front-end technology (like React or Vue.js), back-end framework (like Flask or Django).

Based on the above system architecture and design principles, the automatic evaluation system for English-Chinese interpretation in this study will meet the functional requirements while ensuring high efficiency, availability and security.

5.2 Implementation of Key Modules

In the artificial intelligence-based automatic evaluation system for English-Chinese interpretation, the development of several key modules will help to ensure the effective operation of the system.

(1) Data processing module

Function: Responsible for data collection, re-processing and preparation.

Technical implementation: Data cleaning, formatting, and enhancement using Python.

Key operations: For example, transcribing of audio files, standardization of text, etc.

Table 5 Implementation of Key Modules.

Module	Technology/method	Main operation	Output
Data processing	Python, database system	Data cleaning, formatting, enhancement	The processed data set
Model training	TensorFlow/PyTorch	Algorithm implementation, model training and optimization	Trained model
Evaluation and output	Front-end frameworks (e.g. React)	User interaction, results display	Visual evaluation results

Table 6 Key Features of User Interface and Interaction Design.

Feature	Description	Technology/Tools
Data upload	Users upload interpretation audio and/or text	HTML forms, JavaScript
Progress display	Display data processing and evaluation progress	Ajax, WebSocket
Result display	Present the detailed results of the interpretation assessment	CSS, React/Vue.js
User feedback	Collect user feedback on the evaluation results	Feedback forms, database storage

(2) Model training module

Function: Implement algorithms, train and optimize models.

Technical implementation: Implement LSTM and CNN models using deep learning frameworks such as TensorFlow or PyTorch.

Key operations: these include, model parameter setting, training process management, etc.

(3) Evaluation and output module

Function: Automatically evaluate the interpretation content and provide the result output.

Technical implementation: Build a user interface to display the evaluation results.

Key actions: For example, users submit interpretation samples, and the system displays evaluation scores and feedback.

Table 5 shows an overview of the implementation of the key modules.

Through the implementation of these key modules, the system can efficiently process and evaluate interpretation data, and provide users with accurate and practical evaluation results.

5.3 User Interface and Interaction Design

The user interface (UI) and interaction design serve as the bridge between the user and the system for the AI-based automatic evaluation system of English-Chinese interpretation. An intuitive, easy-to-use interface is crucial for a positive user experience.

(1) Interface layout and design

Main functions: Interface design should include data upload, display of the evaluation progress and display of results.

Design principles: These should be concise and easy to navigate, with an emphasis on user experience.

(2) Interactive process

User actions: These include uploading interpretation samples, viewing evaluation progress and obtaining evaluation results.

Feedback mechanism: The system should give real-time feedback on the processing status such as upload progress, evaluation progress, etc.

(3) Technical implementation:

Front-end technology: Use a modern front-end framework such as React or Vue.js to build the interface.

Back-end integration: Interact with back-end model training and evaluation modules through RESTful apis.

Table 6 shows the key features of user interface and interaction design.

With this user interface and interaction design, the system will provide users with a clear and friendly operating environment. Users can easily upload samples, view evaluation progress and results, and provide feedback on evaluation results.

6. SYSTEM TEST AND EFFECT EVALUATION

6.1 Test Plan and Method

For the automatic evaluation system of English-Chinese interpretation based on artificial intelligence, a comprehensive system test and effect evaluation is the key step applied to ensure its reliability and effectiveness.

(1) Test type

Functional testing: Verify that the various functions of the system work as expected.

Table 7 Key Features of the Test Plan.

Test type	Method	Performance index	Instructions
Functional testing	Automated testing	Success rate	Verify that all functions are working properly.
Performance test	Automated testing	Response time, accuracy	Test the speed and accuracy of data processing by the system.
User experience	User testing	User satisfaction	Evaluate interface friendliness and ease of use.
Safety test	Expert review	Security compliance	Ensure data processing complies with security standards.

Performance testing: Determine the speed and accuracy with which the system processes the data.

User experience testing: Collect user feedback on the friendliness and ease of use of the interface.

Security testing: Ensure that the data processing of the system complies with security and privacy standards.

(2) Test method

Automated testing: Use automated tools for functional and performance testing.

User testing: Invite the target user group to perform the actual operation and collect feedback.

Expert Review: Experts evaluate the overall design and performance of the system.

(3) Performance indicators

Use mathematical models to measure performance, such as response time ($T_{response}$), Accuracy ($Accuracy$), and so on. The following (11) is shown:

$$T_{response} = \frac{1}{N}\sum_{i=1}^{N} t_i \tag{11}$$

(4) Test implementation

Develop a detailed test plan comprising test cases, test data, and expected results.

Perform tests regularly and record the results.

Table 7 shows the key features of the test plan.

Through these comprehensive test plans and methodologies, the system will be thoroughly evaluated to ensure that it meets the expected performance and all safety standards.

6.2 Performance Testing and Analysis

When developing an automatic evaluation system for English-Chinese interpretation, performance testing and analysis is a key step that is used to determine the validity and reliability of the system.

(1) Performance test indicators:

Accuracy (Accuracy)

Response time

Resource Usage: such as CPU and memory usage.

(2) Test data set:

Test using the prepared processed data set.

Includes samples of interpretation of different types and difficulties.

(3) Test implementation:

Perform a series of performance tests on the system, including tests under different conditions (such as different data volumes, different network conditions).

(4) Performance analysis method:

Collect test results, including accuracy, response time, and resource consumption.

Compare the test results under different conditions and analyze the performance changes of the system. This is shown in Figure 2.

The key results of the performance tests are shown in Figure 3.

Through these comprehensive performance tests and analyses, the system's performance can be evaluated under various conditions, potential areas requiring optimization can be identified, and the final performance can be ensured to meet the requirements.

6.3 User Experience and Feedback

When developing an automatic evaluation system for English-Chinese interpretation, it is essential to understand the users' experience and collect their feedback in order to optimize and perfect the system.

(1) User experience evaluation indicators:

User Satisfaction: Obtained through questionnaires and rated on a Likert-type five-point scale ranging from 1 (very dissatisfied) to 5 (very satisfied).

Task Completion Time: The time a user requires to complete a specific task.

Error Rate: How often users experience errors during use.

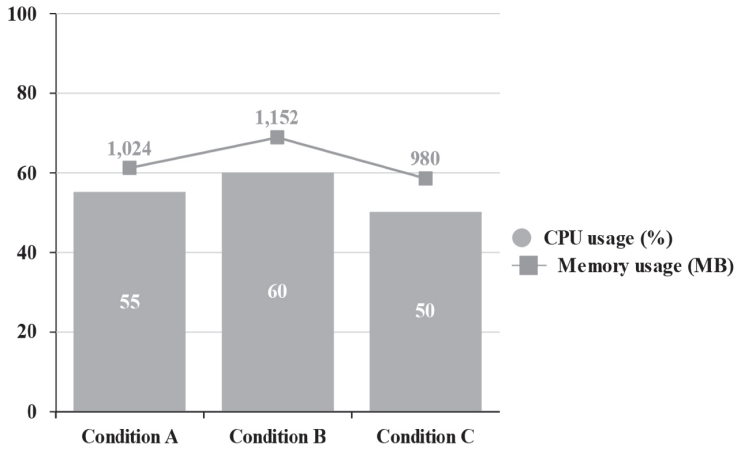

Figure 2 Performance test.

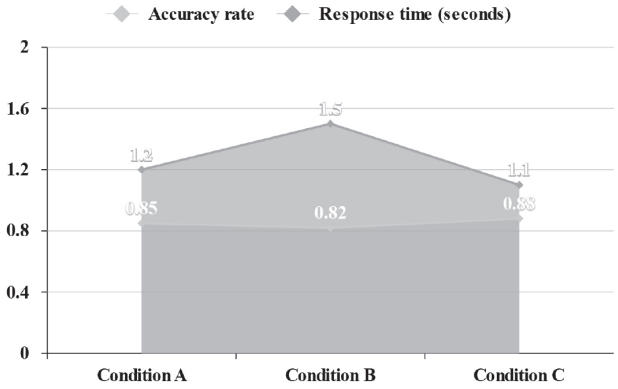

Figure 3 Test results.

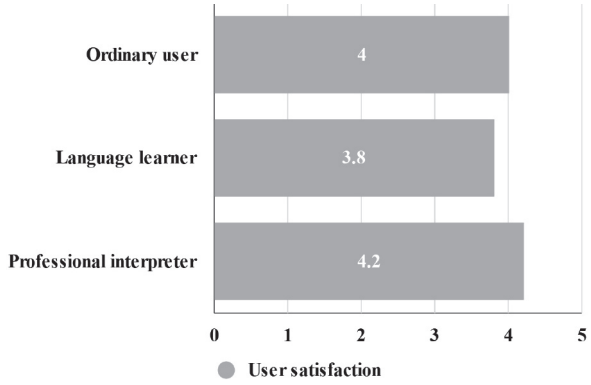

Figure 4 User experience test.

(2) User test group: Users from different backgrounds, including professional interpreters, language learners and ordinary users, were invited to participate in the test.

(3) Feedback collection method:

Online questionnaire survey: Collect users' satisfaction with the system interface and functions.

User interview: In-depth understanding of users' specific needs and experience.

(4) User experience testing and analysis: Implement a series of user tests, record and analyze data, as shown in Figure 4.

Figure 5 shows the key results of the user experience test.

Through comprehensive user experience testing and feedback, we acquired a strong understanding the user's satisfaction and usage of the system, which provide an important basis for further optimization of the system.

7. CONCLUSIONS

This study has successfully developed and implemented an artificial intelligence-based automatic evaluation system for English-Chinese interpreting. The aim is to provide an objective and efficient tool for assessing interpreting quality.

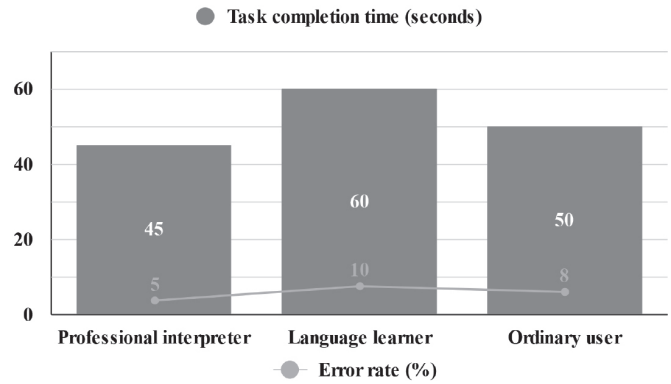

Figure 5 Key results of the test.

Through thorough data collection and teleprocessing, the system ensures the quality and diversity of input data, laying a solid foundation for accurate evaluation. In terms of algorithm design, the innovative approach combines a hybrid model of convolution neural network (CNN) and long short-term memory network (LSTM), performing well in processing interpreting data, especially in dealing with long sentences and complex contexts. The system development follows the principles of molecularity, scalability, and user-friendliness, resulting in a technologically advanced and intuitive final application layer.

When tested and evaluated, the system demonstrated good performance in functional, performance, and user experience tests. User groups with different backgrounds generally provided positive feedback on the ease of use and evaluation accuracy of the system. However, there are some limitations to the study. For example, although efforts were made to ensure data diversity, there are still a limited number of samples of specific contexts or accents. Additionally, the algorithm's accuracy in dealing with extremely complex or low-quality audio needs improvement.

In future work, it is recommended that the datasets be expanded by adding more diverse samples of accents and contexts to improve the model's ability to generalize. Simultaneously, the algorithm should be continuously optimized, especially in regard to speech recognition and semantic understanding, to further enhance the accuracy and reliability of the evaluation. Finally, this study introduces a new technical approach for the assessment of English-Chinese interpretation and provides a valuable reference and foundation for future related research.

8. FUNDING

This projected was supported by the "14th Five-Year Plan" of Shaanxi Education Science 2021, "Research on the Construction of English Interpreting Training System in Applied Undergraduate Colleges under the Background of Language Service" (SGH21Y0451).

REFERENCES

1. Su, W. (2020). Exploring how rubric training influences students' assessment and awareness of interpreting. Language Awareness, 29(2), 178–196. DOI: 10.1080/09658416.2020.1743713.
2. Su, W. (2019). Exploring native English teachers? And native Chinese teachers? Assessment of interpreting. Language and Education, 33(6), 577–594. DOI: 10.1080/09500782.2019.1596121.
3. Irribarra, D. T., & Arneson, A. E. (2023). The challenge of defining and interpreting dimensionality in educational and psychological assessments. Measurement, 221, 113430. DOI: 10.1016/j.measurement.2023.113430.
4. Han, C., Hu, J., & Deng, Y. (2023). Effects of language background and directionality on raters' assessments of spoken-language interpreting. Revista Española de Lingüística Aplicada, 36(2), 556–584. DOI: 10.1075/resla.21009.han.
5. Lu, X. L., & Han, C. (2023). Automatic assessment of spoken-language interpreting based on machine-translation evaluation metrics: A multi-scenario exploratory study. Interpreting, 25(1), 109–143. DOI: 10.1075/intp.00076.lu.
6. Han, C., & Zhao, X. (2021). Accuracy of peer ratings on the quality of spoken-language interpreting. Assessment & Evaluation in Higher Education, 46(8), 1300–1314. DOI: 10.1080/02602938.2020.1855624.
7. Han, C., & Fan, Q. (2020). Using self-assessment as a formative assessment tool in an English-Chinese interpreting course: student views and perceptions of its utility. Perspectives-Studies in Translation Theory and Practice, 28(1), 109–125. DOI: 10.1080/0907676X.2019.1615516.
8. Han, C. (2018). Latent trait modelling of rater accuracy in formative peer assessment of English-Chinese consecutive interpreting. Assessment & Evaluation in Higher Education, 43(6), 979–994. DOI: 10.1080/02602938.2018.1424799.
9. Kim, A. A., Chapman, M., Kondo, A., & Wilmes, C. (2020). Examining the assessment literacy required for interpreting score reports: A focus on educators of K-12 English learners. Language Testing, 37(1), 54–75. DOI: 10.1177/0265532219859881.
10. Fu, R. B. (2018). Metadiscourse and coherence in interpreting. Babel-Revue Internationale De La Traduction-International Journal of Translation, 63(6), 846–860. DOI: 10.1075/babel.00017.ron.
11. Feng, X. (2024). Application of big data in English teaching evaluation and feedback system. Engineering Intelligent Systems, 32(6), 625–634.
12. Hubscher-Davidson, S., & Devaux, J. (2021). Teaching translation and interpreting in virtual environments. Journal of Specialised Translation, (36), 184–192.
13. Yenkimaleki, M., & van Heuven, V. J. (2023). Relative contribution of explicit teaching of segmentals vs. prosody to the quality of consecutive interpreting by Farsi-to-English

interpreting trainees. Interactive Learning Environments, 31(1), 451–467. DOI: 10.1080/10494820.2020.1789673.
14. Cheung, A. K. F. (2020). Interpreters' perceived characteristics and perception of quality in interpreting. Interpreting, 22(1), 35–55. DOI: 10.1075/intp.00033.che.
15. Vertanová, S., & Slobodová, M. (2020). Bilingualism versus monolingualism and its correlation with the interpreting mode. Circulo De Linguistica Aplicada A La Comunicacion, (84), 167–173. DOI: 10.5209/clac.72003.
16. Yan, J. X., & Luo, K. T. (2023). Audio description and interpreting training: a comparison of assessment criteria from the perspective of learners. Perspectives-Studies in Translation Theory and Practice, 31(1), 451–467. DOI: 10.1080/0907676X.2023.2186794.
17. de la Fuente, E. A., Fuertes, R. F., & García, O. A. (2019). Bilingual children as interpreters in everyday life: how natural interpreting reinforces minority languages. Journal of Multilingual and Multicultural Development, 40(4), 338–355. DOI: 10.1080/01434632.2018.1518985.
18. Warnicke, C., & Plejert, C. (2018). The headset as an interactional resource in a video relay interpreting (VRI) setting. Interpreting, 20(2), 285–308. DOI: 10.1075/intp.00013.war.
19. Vranjes, J., & Brône, G. (2021). Interpreters as laminated speakers: Gaze and gesture as interpersonal deixis in consecutive dialogue interpreting. Journal of Pragmatics, 181, 83–99. DOI: 10.1016/j.pragma.2021.05.008.
20. Hijazo-Gascón, A. (2019). Translating accurately or sounding natural? The interpreters' challenges due to semantic typology and the interpreting process. Pragmatics and Society, 10(1), 72–94. DOI: 10.1075/ps.00016.hij.
21. Zhong, Y., Xie, J. C., & Zhang, T. (2021). Crowd creation and learning of multimedia content: an action research project to create Curriculum 2.0 translation and interpreting courses. Interpreter and Translator Trainer, 15(1), 85–101. DOI: 10.1080/1750399X.2021.1880695.