Intelligent Chinese Speech Learning Correction System

1st Keying ZHU
School of Information and Electronic Engineering
Zhejiang University of Science & Technology
Hangzhou, China
z_keying1004@163.com

2nd Yuefeng CEN
School of Information and Electronic Engineering
Zhejiang University of Science & Technology
Hangzhou, China
cyf@zust.edu.cn

3rd Jiaming GU
School of Information and Electronic Engineering
Zhejiang University of Science & Technology
Hangzhou, China
gujm1004@163.com

***Abstract*—In order to meet the learning needs of Chinese learners and help them correct the nonstandard pronunciation of Mandarin, the project team applies intelligent speech technology to Chinese learning, designs and develops a intelligent Chinese speech learning correction system. The system realizes Chinese learning and speech correction for all users through the combination of intelligent mobile terminal and Web background management terminal. Users improve their oral communication skills in Chinese through intensive listening and follow-up reading, correct pronunciation errors in spoken language with the help of the system's audio-video dual-modal algorithm, improve their understanding of Chinese characters through writing training, and practice Chinese ability in community communication. After the application test, the system can effectively improve the learners' Chinese oral expression ability.**

Keywords—Chinese learning, assessment, speech correction

I. INTRODUCTION

At present, China has established 541 Confucius Institutes and 1,170 Confucius Classrooms in 162 countries or regions. A total of 75 countries around the world have incorporated Chinese into their national education systems. More than 4,000 foreign universities have offered Chinese courses, and they are learning Chinese outside of China. The number of people is about 25 million, and it is gradually increasing with a good trend [1]. From 2016 to 2020, more than 40 million people took Chinese proficiency tests(HSK) and YCT which is Chinese Test for Primary and Secondary Schools[2]. thus it can be seen that more and more people are learning Mandarin Chinese internationally, and at the same time, the demand for excellent Mandarin teachers abroad is also showing a significant growth trend. The penetration and influence of information technology on education has become increasingly prominent, So the integration of Chinese learning and information technology has become one of the important driving forces for the reform of Chinese education [3]. Since the beginning of the 21st century, with the development of technologies such as 5G, artificial intelligence, big data, and cloud computing, smart education empowered by technology has gradually come to fruition from the previous concept.

Under the new situation of normalization of online teaching, international Chinese education must adapt to the needs of the intelligent era, and use intelligent technology to reform the teaching environment, teaching mode, teaching content, teaching organization, teaching evaluation, etc., to provide learners with more personalized and Accurate Chinese wisdom education can solve the major problem that the sustainable development of international Chinese education cannot be achieved by offline communication at home and abroad [4]. In China, due to the heavier local accents in some regions, the Mandarin of Chinese people is not standard. In the process of interpersonal communication in daily life, people usually have many misunderstandings due to non-standard pronunciation, which deepens the difficulty of communication between the two parties, and also brings a lot of inconvenience in life. How to effectively improve the quality of Chinese people's Mandarin has become a major difficulty in people's thinking [5]. To this end, under the background of Chinese smart education at home and abroad, based on the computer-aided environment to improve the accuracy of learners' pronunciation, an intelligent Chinese phonetic learning system is built to assist Chinese teachers, help domestic people to improve their Mandarin level, and help foreign friends. Learning Mandarin Chinese also plays a role in deepening the feelings of friends at home and abroad. Compared with software such as Yinshu, Rehabilitation Cloud, and Dr. Qiyin, which require payment, this system is free for learners and has the function of speech correction.

II. THE RESEARCH AND DESIGN OF THE SYSTEM

A. the Proposal of the System

The project team found that the global educational mobility continued to weaken, while the number of people learning Mandarin Chinese continued to grow, which greatly affected the learning of Chinese. In the actual spoken Mandarin teaching in China, everyone's learning habits are different, and the materials for oral Chinese learning are scattered, while the traditional online oral language learning software has a very high personal abandonment rate due to the boring and lonely learning process.

Based on the current background of the times, the project team starts with the needs of domestic and foreign friends for oral Mandarin learning, alleviates the social status quo of the serious lack of online Mandarin personalized education, and fills the gap in the online Chinese phonetic learning correction software on the market, so that Mandarin learners do not have to Get high-quality spoken language learning and pronunciation error correction while worrying about cost [6]. To this end, the project team proposed and designed and developed the system, which is a new dual-modal intelligent Chinese speech correction platform developed under the background of the current shortage of Chinese speech learning and correction resources in the market. Targeting at home and abroad Chinese learners and different learning scenarios, it helps to more comprehensively implement the core concept of " internationalization at home ".

B. the Design of the System

This system is mainly aimed at Chinese people with non-standard Mandarin and foreign friends who have Chinese learning needs, helping domestic and foreign learners to learn spoken Chinese efficiently. The overall structure of the system is shown in Fig. 1. It is mainly divided into modules such as intensive listening, follow-up reading, speech correction, learning assessment, writing training, book reading, and community communication. Each module is independent and complementary to each other, ensuring the stable operation of the system and the hierarchical learning of users. In the wave of " Mandarin fever ", the system builds up the application of Chinese learning and speech correction through the Internet. In order to meet the user ' s learning of Chinese spoken language, the project team constructs the intelligent Chinese speech learning correction system.

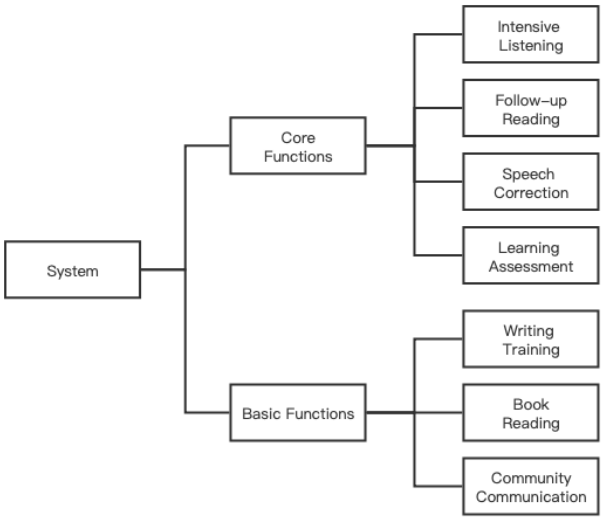

Fig. 1. The overall structure of the system.

This system assists teachers in teaching by "seeing, listening, practicing, commenting, writing, and discussing", and provides users with a new channel to quickly improve their oral Chinese learning. This system concretizes the way of oral language learning into three core learning functions: intensive listening, follow-up reading and speech correction. Each function has a corresponding scoring mechanism, and finally a comprehensive learning evaluation score is obtained based on the three scoring results, allowing users to clarify their own Chinese speaking learning level. In terms of basic functions, users can exercise their writing ability through the writing training, and at the same time, they can watch the books they are interested in through the book reading. The system also provides a community communication function to help users consolidate and strengthen the practical application ability of spoken Chinese.

C. Each Functional Module

The following is a detailed description of the functions of each module of the system and its architecture.

1) Intensive Listening: The system collects a large number of high-quality movie and video clips, which are classified according to types such as suspense, history, military, romance, science fiction, etc., which enriches users' application requirements for spoken Chinese in various scenarios and facilitates users to search for oral Chinese learning materials. Users can choose their favorite categories and then select videos for listening practice. Users can also listen carefully through browsing history and favorites. Users can listen to Chinese carefully by choosing to open subtitles and close subtitles, which can gradually strengthen their learning of spoken Chinese. After listening to Chinese intensively, users need to rate their own listening, and give their own clear assessment of this listening exercise, which is helpful for users to improve their Mandarin level in the process of self-evaluation and comparison. This self-rating result It also serves as the basis for the system to evaluate the user's comprehensive learning.

2) Follow-up Reading: While improving Chinese listening ability, users can use the follow-up reading function to perform follow-up reading according to the high-quality movie and video clips provided by the system, so as to improve their oral Chinese communication ability. The system only records the user's audio under the follow-up reading function. The user can choose to follow the sentence by sentence or read the whole article, and after the system records the user's follow-up voice, the user can compare it with the original audio to find out his own shortcomings. At the same time, the system will mark the text with different pronunciation from the original audio in red to attract users' attention to the text with non-standard pronunciation. The user can repeat the follow-up exercise. After the user has recorded the entire follow-up reading, the system will give the score based on the last follow-up reading result.

3) Speech Correction: Under the follow-up reading function, the film and video clips that the user has completed follow-up will appear under the speech correction module. The speech correction function can help users correct the pronunciation of sentences that have always been non-standard during follow-up reading. Users need to record their own audio and video under the speech correction function. Under the movie video clips, the user reads the sentences marked in red in turn, and the system first determines whether they are accurate or not. If some of the text is still marked in red, the system will cut and analyze the audio and video during the just follow-up process according to the audio-video dual-modal algorithm. According to the video mode, it identifies and corrects the movement details such as the pronunciation action of the lips and the zooming position of the front and rear nasal sounds. According to the audio mode, it identifies and corrects the sound problems such as the front and rear nasal sounds and tonal pronunciation of the content read aloud by the user. Finally, the results of the cutting analysis are fused and summarized. The system will give a spoken pronunciation video of word-by-word pronunciation and a text description suggestion for the pronunciation of the word, helping users to more accurately correct the spoken pronunciation by watching the video and reading the text. The user can repeat the follow-up correction repeatedly. After the user's last follow-up recording and correction is completed, the system will give a score for improving the follow-up accuracy through voice correction. For example, if the score is 80 points after the follow-up reading, and the score is 90 points after correction by the speech correction function, the speech correction score of the movie video clip is (90-80)/(100-80))*100=50 points.

4) Learning Assessment: The user's comprehensive learning situation is presented in scores, ranging from 0 to 100. The system averages the user's self-score after each intensive listening, and the average value will be used as the user's intensive listening score. Similarly, the system takes an

average of the scores of each follow-up reading of the user, and the average value will be used as the score of the user's follow-up reading. The system also averages the scores of each speech correction of the user, and the average value will be used as the score of the user's speech correction. However, if a certain movie video segment under the speech correction module is not full score, and the correction has not started, it will be processed with 0 points; Finally, the scores of intensive listening, follow-up reading and speech correction are weighted and averaged according to the weights of 0.2, 0.35 and 0.45 respectively, and this value is used as the user's current learning evaluation result.

5) Writing Training: Chinese characters have the characteristics of being figurative, and only a better understanding of the characters can have a better writing training effect. After the learner has determined the writing content of interest, the system provides the user with a relevant video containing the text, including other relevant introductions such as the origin and evolution of the text. After watching the video, users will have a deeper understanding and mastery of the text to be written. The user takes a picture of the written Chinese character and uploads it, and the system provides accurate and effective Chinese character modification suggestions and scores for the user's writing result through the feature extraction algorithm and semantic analysis algorithm of artificial intelligence.

6) Book Reading: While learning spoken Chinese, users can use the book reading function to watch books they are interested in, and add books to the bookshelf to read books that have not been read or have no time to read for the time being, making it easier for users to find the book. Users can choose to read the book or listen to the book to read the book. If they encounter difficult reading, they can also click to query. The text that can be queried usually appears in the form of phrase pairs, and the user can know the pronunciation of the word and its Chinese meaning. The book reading function further strengthens the user's mastery of spoken language, deepens the user's understanding of Chinese, and sublimates the user's love for Chinese culture.

7) Community Communication: Users can communicate in the community, which is open to all registered users. Users can share their learning experience of spoken Chinese and recent life anecdotes in the community by sending text, pictures and short videos. Users can search for interesting content through the search bar. Based on the collaborative filtering algorithm, the system pushes high-rated dynamic content that users are interested in according to their preferences. Users can like, comment, favorite and rate the updates sent by others, and users can also see other users' comments and like the comments under the same update. Users can start chatting by clicking on their avatars. There are various chat methods. Users can choose text input, or send short voice messages. They can also send emoticons, photos and videos to strengthen their communication skills through practical exercises. The user can also follow the person he likes, and the user can conveniently and timely receive the immediate dynamic of the person he has followed through the follow. It is conducive to establishing friendly communication between users, narrowing the distance between people who learn spoken Chinese at home and abroad, and enhancing the feelings between international friends.

III. TECHNOLOGY ROUTE

A. Technology Framework

In order to make the system available to all users who need to learn spoken Chinese, the project team adopts both iOS and Android development, and also develops a background management website as the management terminal. The system is developed using the Flutter framework and introduces Redux-middlewar to handle more complex business [7]. Realize a set of code, multi-end generation, and realize complex and flexible interface design through composable space combination and rich animation library.

The client uses Dart as the programming language for development. Dart is one of the few languages that supports both JIT (just-in-time compilation) and AOT (compile-before-running), so it also shows the characteristics of fast running speed and good execution performance during the development and running process.

In server-side development, we use SpringBoot as the framework, MVC as the main architecture, and MyBatis as the persistence layer framework. In terms of performance optimization, we introduce Kafka as a message middleware to process message queues in the state of large amounts of data, and use it as push processing for streaming media and learner forums. In addition, we use Redis distributed caching technology to safely back up user data, and also widely used in operations such as displaying, deleting and filtering lists [8].. Finally, we use Nginx-based Ip-Hash strategy to bind clients and servers to achieve one-to-one, and optimize Nginx's reverse proxy and load balancing .

SpringBoot combines MVC architecture as a backend framework This model is used for layered development of programs. We use the SpringBoot framework to greatly facilitate the plug-in deployment process and simplify the initial construction and development process of the application. Using the Controller as a proxy for the connection, the View and Model are easily connected, allowing multiple Views to share a Model.

MyBatis is a semi-automatic persistence layer framework that encapsulates the process of JDBC operating the database. We only need to focus on SQL itself, which uses simple XML and annotations for configuration and original mapping, allowing us to use the use of object oriented programming method (OOP) to operate the database.

Redis is an in-memory database, which not only improves user access speed, but also saves data space and reduces database pressure while ensuring data integrity in the process of data caching. We use its own hash data structure to realize the data storage of the learning forum, and finally use the master-slave data backup mode to cache messages, set the expiration time according to the key, and automatically delete it to improve the security of user data backup.

Kafka is a distributed, partition-supported messaging system that uses its batch processing mechanism to meet the needs of large data volume transmission processing and low latency.

The Ip Hash strategy is a built-in strategy. The front-end access IP is hashed first, and then requests are allocated to different back-end nodes according to the results. Through Nginx implementation, each front-end access IP will access a

back-end node fixedly, which can avoid the problem of session sharing.

B. Core Technology

1) Speech Correction Technology: The user records audio and video through the interactive interface, and inputs the audio and video into the system, the system will automatically convert the video in mp4 format into audio part and video part, and the video part will be automatically decoded and cut into a picture sequence, one frame by one picture. The system chooses 24 frames as the unified input metric of the model to achieve two balanced effects of no loss of information and as small a model input as possible. In order to improve the robustness of the model, and considering that users speak at different times on each word, the system uses multiple copying operations starting at random starting points to replace the method of re-copying at fixed positions. At the same time, the system adds the processing methods of random frame loss and random movement of pictures. Random frame loss occurs when the number of samples is greater than a specified specific value, and the lips of a series of pictures are not in the same position, so as to enhance the generalization ability of the model.

In terms of feature extraction, in terms of video, the face-alignment toolkit of the dlib library in Python is used to extract the face feature points of the face part of the video picture sequence. In practical situations, it is difficult to predict the shape of the face by using features with high confidence, and to ensure the accuracy of the features extracted from the shape of the face [9]. To solve this problem, the system uses an iterative approximation method to extract the sample features based on the current predicted shape, and then uses the extracted features to update the vector when predicting the shape. This process is repeated several times until convergence. Finally, the corresponding 20 lip key point positions are obtained through the lip key point number, and the relative radian between the key points is used as a feature [10]. Subsequently, all the features of the entire video sequence are combined into a 2D tensor.

In terms of audio, the user's input is evaluated by using the characteristics of different frequencies, and the two feature tensors are respectively sent to the bidirectional long-short-term memory network Bi-LSTM for judgment. The overlapping segments of the two sequences have lower confidence. That is, the mispronunciation fragment. After the audio data entered by the user is processed by audio segmentation, the audio of a single word is sent to the trained lstm model for evaluation and scoring.

Based on the large-scale Chinese data set dedicated to the pronunciation of cloud port language, the system can effectively train and establish a high-accuracy spoken language recognition model. Finally, the two features of audio and video are input into the multi-modal deep neural network. Finally, the system performs comprehensive evaluation and scoring according to the scores obtained from video processing and audio, and provides users with accurate pronunciation of each text with non-standard pronunciation according to the output results of the network. Video suggestions and text suggestions for speaking improvement.

2) Writing Recognition Technology: Based on the current mainstream image text recognition model, a text recognition model based on attention mechanism is proposed, and its framework is shown in Fig. 2.

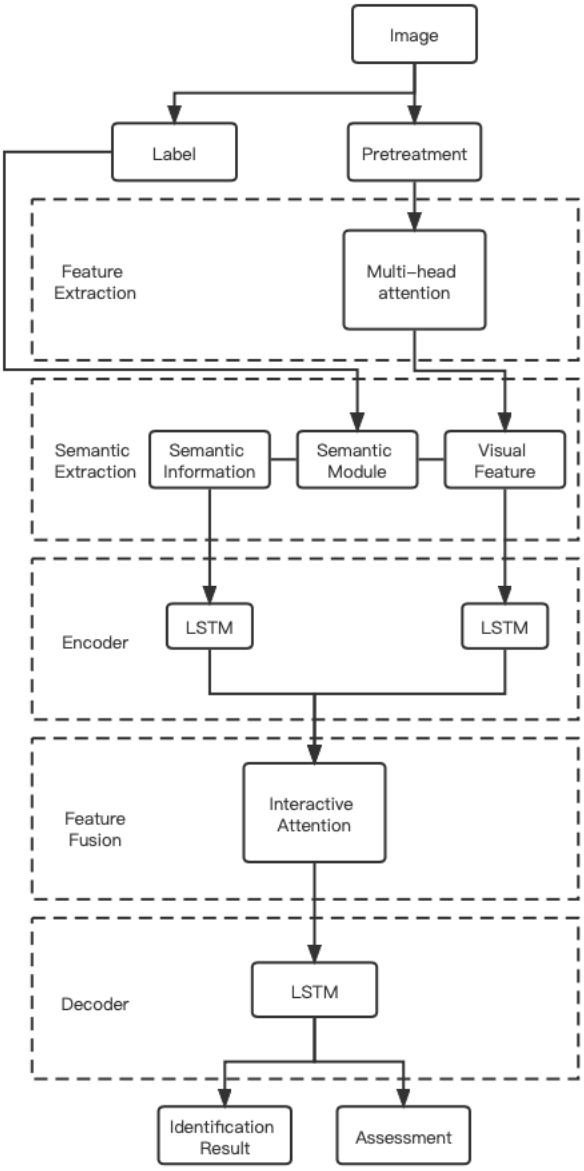

Fig. 2. Text recognition model framework based on attention mechanism.

The user uploads the written text in the form of a photo. After the image is uploaded, the system preprocesses it to eliminate irrelevant information as much as possible, highlight useful real information, and then perform grayscale processing.

The model is mainly composed of five functional modules: feature extraction, semantic extraction, encoder, feature fusion and decoder. The feature extraction module uses the multi-head attention mechanism to extract the deep visual feature information of the image. In the specific implementation, the number of heads is set to 8. Q, K, and V are fixed single values and are input to 8 Linear layers, so as to "focus" on the input values from 8 angles. The obtained 8 single-head attention features are used for feature fusion, and then the multi-head attention value of the input information is obtained through the full-connection neural network for advanced conversion, so as to realize the deep feature extraction of the input information. In order to make up for the fact that the current image text recognition model ignores the semantic information of the text to cause text recognition errors, the

semantic information features of the image text are extracted in the image text recognition to strengthen the role of the semantic information in the image text recognition. Then, the visual feature information and semantic feature information are encoded through a two-layer bidirectional Long Short Term Memory Networks (LSTM). In order to effectively fuse image visual information with text semantic information and retrograde, a feature fusion method based on interactive attention mechanism is adopted, that is, visual features interact with semantics to obtain image recognition features that are more critical and accurate for the current recognition task. Achieve efficient fusion of visual features and semantic features. Finally, a bidirectional long-short-term memory neural network is used to decode the fused features to generate text recognition results.

IV. THE IMPLEMENTATION AND HIGHLIGHTS OF THE SYSTEM

A. *the Implementation of the System*

After the completion of the development of the system, each sub-function can run normally. The home page nterface and speech correction interface are shown in Fig. 3. Through the practical application of the system by the surrounding students, the results show that the users can effectively improve the spoken Chinese expression ability.

Fig. 3. The home page and speech correction.

B. *the Highlights of the System*

This system provides a six-in-one all-round teaching method for Chinese teaching, which greatly meets the needs of all kinds of Chinese learners at home and abroad. The speech correction function helps learners to separate and decode the input audio and video, conduct in-depth analysis of lip movements and pronunciation, and finally give more accurate suggestions for improving the spoken language of the user's non-standard pronunciation. Compared with the blurred text image, the writing training function can also more accurately identify the text and image features uploaded by the learners, and efficiently integrate the visual features and semantic features to give the recognition results and evaluation results. Learners in different regions can also conduct real-time cross-regional real-time communication through community communication, improve oral language and improve interpersonal communication skills, and shorten the distance between each other.

V. CONCLUSION

The project team designed and implemented the intelligent Chinese speech learning correction system. The system combines intelligent speech technology and Chinese language learning to provide solutions for problems such as non-standard Mandarin pronunciation and high demand and limited resources for Chinese learning. The system has been tested on a small scale in the university community, and has achieved good results. Some teachers have clearly proposed that the design concept of the system meets the needs of Chinese learning in the current era and promotes the implementation of international Chinese education.

ACKNOWLEDGMENT

This research is supported by the general planning fund for Humanities and Social Sciences Research of the Ministry of Education (No.17YJA880004) in 2017.

REFERENCES

[1] NA-NA NIU and DONG-WEI LI, "Take the Confucius Institute of Traditional Chinese Medicine at the University of Pecs as an Example," *Journal of North China University of Science and Technology(Social Science Edition)*, vol. 21, pp. 147-741, July 2021.

[2] HUANG, F. L, Wu and J. Y, "Study and implementation of intelligent e-learning system for modern spoken Chinese," *International Conference on Machine Learning and Cybernetics*, Vol. 5, pp. 2968-2974, 2009.

[3] HO and W. Y. J, "Coming here you should speak Chinese," the multimodal construction of interculturality in YouTube videos, *Language and Intercultural Communication*, 1-19, 2022.

[4] ZHAN, H., CHENG and H. J, "The role of technology in teaching and learning Chinese characters," *International Journal of Technology in Teaching and Learning, 10(2), 147*, 2014.

[5] QINMING, C. and Jun L., "Analysis of Teaching Treatment Strategy of Foreign Students Majoring in Overseas Chinese Education," *International Conference on Education Technology and Economic Management (ICETEM 2019)*, pp. 806-811, 2019.

[6] YANG SU, "A Study of Curriculum Learning Experience and influence Factors od International Doctoral Students Studying in China under the BackGround of 'the Belt and Road'", *International and Comparative Education*, 9rd, pp. 8-26+35, 2019.

[7] HUANG, H., "Design and Implementation of a College English Listening Learning System Based on Android Platform," International Journal of Emerging Technologies in Learning, 13(7), 2018.

[8] ZHENG, Z., CHENG, J., and Peng, J., "Design and implementation of teaching system for mobile cross-platform," *International Journal of Multimedia and Ubiquitous Engineering*, *10*(2), 287-296, 2015.

[9] LAN, Y., Theobald, B. J., Harvey, R., Ong, E. J., and Bowden, R, "Improving visual features for lip-reading," *In Auditory-Visual Speech Processing*, 2010.

[10] Adeel, A., Gogate, M., Hussain, A., and Whitmer, W. M., "Lip-reading driven deep learning approach for speech enhancement," *IEEE Transactions on Emerging Topics in Computational Intelligence*, 5(3), 481-490, 2019.

[11] Rodriguez, P., Velazquez, D., Cucurull, G., Gonfaus, J. M., Roca, F. X., and Gonzalez, J., "Pay attention to the activations: a modular attention mechanism for fine-grained image recognition," *IEEE Transactions on Multimedia*, 22(2), 502-514, 2019.