Hybrid Japanese Language Teaching Aid System with Multi-Source Information Fusion Mapping

Rui Zhang

School of Foreign Studies, Lingnan Normal University, Guangdong, Zhanjiang 524048, China

Correspondence should be addressed to Rui Zhang; zhangr@lingnan.edu.cn

Received 4 July 2022; Accepted 22 August 2022; Published 27 September 2022

Academic Editor: Lianhui Li

Copyright © 2022 Rui Zhang. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Learning Japanese can enhance competitiveness in a globalized economy, and we address the problems of poor open-source Japanese language teaching, cumbersome teaching tasks, and a single teaching model. We propose a hybrid Japanese teaching aid system with multiple information fusion mapping, which can effectively improve the efficiency of Japanese teaching and reduce the tedious human teaching procedures. The system is divided into two branches of Japanese language recognition, namely, the Japanese text recognition branch and the Japanese voice sequence recognition branch. In the Japanese text recognition branch, we integrate attention mechanisms and long short-term memory networks as the basic network for Japanese character text recognition. In addition, we set up separate text feature recognition systems for Japanese computer writing and handwriting to prevent feature overlap problems. For Japanese voice sequence recognition, we used a combination of memory gating unit and encoder, based on the network still extending the structure of the deep neural network and using the residual structure connection in the gating unit to avoid the gradient disappearance problem. At the end of the system, we use a softmax layer to connect the text recognition and voice recognition networks to form a Japanese language teaching aid system. To verify the efficiency of our system, we selected the Japanese text recognition public dataset and voice recognition public dataset for experimental validation. To match the practical application of the system, we created our dataset based on the dataset standard and conducted experimental validation. To compare other Japanese recognition methods, we selected the six most representative Japanese recognition algorithms for experimental comparison. To ensure the balance of the experiments, each algorithm is trained in a separate experimental environment for modeling and tuning parameters. Experimental performance and the experimental results show that our method significantly outperforms the other methods and has better system stability.

1. Introduction

Economic globalization has become an inevitable trend, and a tool to assist it is linguistic communication between different countries. Learning a foreign language can enhance the competitiveness of economic globalization and contribute positively to the cause of multiculturalism. English is now the dominant language in globalization, but for people in non-English-speaking regions, small languages can enhance the efficiency of economic development in non-English-speaking countries. Japan has played an irreplaceable role in today’s economic globalization process, and Japanese, as one of the minor languages, is not very difficult to learn and master. Also, Japanese is one of the international languages highly valued by the Malaysian education sector [1].

With the full coverage of the 5 G communication network, Japanese language teaching is gradually shifting from traditional face-to-face teaching to online learning and remote instruction mode.

The innovation of small language teaching methods is the inevitable result of technological development. Japanese language teaching has changed from the digital media technology model to the Internet information technology model, different computer technologies have given Japanese language teaching a new teaching model, and students’ learning experience and learning efficiency are higher. The Japanese government is now actively contacting Japanese-friendly countries to jointly build a new field of Japanese language teaching based on WEB. In the traditional Japanese language teaching model, Japanese language teachers are
burdened with tedious teaching tasks. Japanese language intelligent teaching system can reduce the teaching tasks of Japanese teachers and also help students to strengthen the foundation of the Japanese language [2]. The Japanese language learning environment is also one of the most influential factors in Japanese language learning, and an excellent Japanese language learning environment can help students quickly improve their Japanese speaking and memorization skills. The Japanese learning environment in an offline classroom is heavily dependent on the teaching style of the Japanese teacher, but the Japanese learning environment in the Japanese smart teaching model is pre-set according to the student’s Japanese ability and is more friendly to different students. The Japanese smart teaching system includes more than just Japanese grammar and word teaching tasks; according to the latest Japanese smart teaching research, researchers are creating a virtual environment for Japanese language learning, and virtual reality and augmented reality technologies are increasingly being used in the Japanese smart learning system. This not only enhances the fun of Japanese lessons but also boosts student motivation [3].

Some researchers who focus on building Japanese intelligent teaching systems have found that interactive systems increase Japanese language perception in Japanese language learning, so most Japanese intelligent teaching models use interactive learning methods. Researchers have transplanted visual sensing technology to assist in Japanese language teaching by analyzing student-teacher interactions to recommend appropriate Japanese learning methods for students and more relevant Japanese teaching programs for teachers after class. Some researchers have embedded voice sensors into spoken Japanese-assisted learning to process and analyze data on students’ spoken Japanese and provide students with word pronunciation correction and grammar optimization suggestions, and students can receive a real-time Japanese pronunciation suggestion after each Japanese pronunciation. Long-term Japanese speaking practice and feedback can leave data records in the Japanese Speaking Assistance System, which will analyze big data on individual students’ pronunciation habits and speaking error point to provide each student with an adapted Japanese speaking training plan [4]. With the change in information technology, the open-source nature of portable electronic device systems has increased. Researchers aim to increase the open-source nature of intelligent Japanese language teaching systems to help students use their free time to learn Japanese efficiently. Researchers have developed Japanese smart teaching systems that integrate WEB and APP, allowing students to select online lessons, learn vocabulary and grammar, practice speaking, learn Japanese culture, view Japanese news, etc., on various portable devices. Some researchers have conducted research on Japanese text recognition to detect handwritten Japanese, improving students’ handwritten Japanese skills, and enhancing the handwritten Japanese experience in Japanese classes [5].

In response to the problems of poor open-source Japanese language teaching, cumbersome teaching tasks, and a single teaching model, we propose a hybrid Japanese teaching aid system with multiple information fusion mapping that can effectively improve the efficiency of Japanese language teaching and reduce the cumbersome human teaching procedures. The system is divided into two branches of Japanese language recognition, the Japanese text recognition branch, and the Japanese voice sequence recognition branch. To verify the efficiency of our system, we selected the Japanese text recognition public dataset and voice recognition public dataset for experimental validation. To match the practical application of the system, we make our dataset according to the dataset standard and perform experimental validation.

The rest of the paper is arranged as follows. Section 2 describes the work related to Japanese voice recognition and text recognition. Section 3 describes in detail the principles and implementation process related to Japanese language recognition methods. Section 4 shows the related experiment setups, experimental dataset, and analysis of experimental results. Finally, Section 5 summarizes our study and reveals some further research work.

2. Related Work

The Japanese language system contains a large number of Chinese characters, so the Japanese text recognition methods have a lot in common with the Chinese language recognition methods. Some researchers studying Japanese document recognition use layout analysis [6] to segment Japanese fragments and then use fixed pixel frames to extract pixel information from Japanese fragments in an iterative manner [7]. The extracted features can be categorized into Japanese character feature databases based on manual labels, and through the neural network layer, different Japanese character features will be linked in independent mappings based on the neural network. For Western languages, most scripts are composed of Arabic letters, while the Japanese system is composed of hiragana, katakana, and kanji. In the character segmentation work, the Japanese character segmentation work is completely different from the English system. In the literature [8–10], the method of segmentation followed by merging was proposed in the study. Due to the high workload of Japanese character segmentation, the experimental cost is high and it is easy to make segmentation errors. Therefore, in the work of Japanese character segmentation, researchers divided it into written Japanese segmentation and handwritten Japanese segmentation. The written style is more standardized, and the character segmentation work is easier, while the handwritten style varies from person to person and the character segmentation is more difficult. The accuracy of character segmentation directly affects the performance of the whole Japanese recognition system. The early research on Japanese character segmentation used machine learning algorithms as the main character feature learning method, and later researchers introduced deep learning methods into character research, which greatly improved the efficiency of hiragana and kanji character segmentation.

The first deep neural network applied to Japanese intelligent teaching system can significantly improve the
recognition accuracy of the Japanese language. The accuracy of support vector machines in character segmentation cannot meet the subsequent Japanese language processing, so many researchers try to use deep neural networks instead. Researchers in the literature [11] have proposed a dual-linked neural network framework by fusing them convolutional neural networks and long short-term memory units. The method aims to improve the recognition accuracy of Japanese computer writing style and concludes by proposing an association with Japanese handwriting style, which provides a great reference for later Japanese handwriting style recognition. Researchers in the literature [12] analyzed the current problems faced by Japanese text recognition work and proposed a handwriting grading algorithm based on the difficulty of recognizing Japanese handwriting corpora. The study in the literature [14] addresses the problem of offline Japanese language recognition by proposing a framework for the fusion of two-layer long short-term memory units with temporal classification algorithms. The literature [15] investigated the relationship between Arabic script recognition and Japanese language recognition, and successfully transposed the Arabic language recognition model to Japanese language recognition research, and experiments proved that the method achieved effective results. The studies in the literature [16, 17] are end-to-end training models, and the method embeds the model into the Japanese language intelligent recognition system by pretraining the model, which solves the compatibility problem between the model and the system and reduces the computational cost.

Japanese voice recognition belongs to natural language audio processing, and for Japanese voice recognition, it is first necessary to convert the voice signal into a linguistic feature vector. Then, the Japanese voice features are enhanced by output processing by simulating human ear perception features. Then, the mapping of the voice signal to Japanese text features is completed by matching voice signal features with Japanese features through linear prediction and perceptual prediction. There are also a large number of research results in the field of voice recognition. The research in literature [18] has broken new ground in the field of voice recognition. The authors proposed a voice sequence matching model based on dynamic time regularization, which is simple to understand and has a high recognition correct rate, but is computationally intensive and requires high hardware equipment. So far, the method is still used in voice recognition of access control systems. The literature [19] improves on the former, optimizes the recognition accuracy of small vocabulary and isolated words in voice recognition systems, and also proposes the concept of frequency scale recognition to improve the generalization of voice recognition systems. The literature [20] proposed a voice recognition model based on vector quantization with a sub-parameter model, which requires less computer memory and has better recognition resulting in large segment voice decomposition. Researchers in the literature [21] proposed a segmented fuzzy clustering algorithm to visualize voice sequences and use vector quantization errors to replace the output probabilities of hidden Markov models, and the network model was experimentally shown to have good performance in voice recognition. Researchers in the literature [22] proposed a fusion model of the hidden Markov model and a self-organizing neural network to obtain preceding parameters by analyzing filter sets in voice signals and then using a self-organizing neural network to predict the mapping relationship between voice and text. The experimental results prove that the model has good robustness and stability.

3. Method

3.1. Hiragana and Katakana Feature Classification. The Japanese language system consists of hiragana and katakana, and the kanji part can also be represented by hiragana and katakana; therefore, the characteristic classification of hiragana and katakana has a great influence on the Japanese recognition system. Researchers in the literature [23] have proposed coarse and fine classification systems in the study of feature classification, and both systems use a combination of line segments and dots. For different hiragana and katakana contours, researchers have purposely designed different stroke contour recognizers. Some researchers have tried to use Markov random field algorithms with unstructured features as the main baseline for contour recognition [24]. For Japanese handwriting, this method cannot accurately obtain the mapping relationship between handwriting patterns and the standard Japanese system, and the structural information is easily lost at the temporal level. The literature [25] modified the recognition sequences of Japanese and Chinese characters based on the former to enhance the acquisition of structural and nonstructural features for the Japanese recognition algorithm. We synthesize the former study and propose a mosaic classification method with coarse classification and fine classification. In our classifier, we set up Markov random field structure classifier (MRF-C), hidden Markov structure classifier (HMM-C), and quadratic discriminant function classifier (QDF-C). The hiragana feature classifier we designed is shown in Figure 1.

For the structure recognizer, we used a top-to-bottom order of hiragana contour trajectory extraction. We reconstruct the hiragana trajectory features by taking the character trajectory start point as a unary feature point, and using the unary feature point as the center, we diverge to adjacent coordinate differences as binary features. The binary features are input to the Markov random field model as nodes, and the coarse classifier will preferentially generate large probability category labels, and after the first matching of large probability category labels is completed, the
character feature vectors with finer dimensions are obtained after inputting the monadic features and binary features to the fine classifier. The hidden Markov model cannot complete the binary feature of point-to-point trajectory recognition and classification. Therefore, single-point hiragana character features are graded by random fields to complete feature traversal.

3.2. Attention Mechanism Feature Extraction. We divided the Japanese computer writing style and handwriting style into two separate branches, and we designed specialized attention mechanisms and long short-term memory networks to decompose the different hiragana contour features. For the Japanese computer-written style, we mainly designed unstructured fine feature recognizers. Based on previous studies, Japanese hiragana characters need to be converted to 2D RGB images for storage, and then stroke features are extracted based on predefined handwriting directions. Some researchers tried to normalize the histogram method for hiragana strings, but the results were not very good. Some researchers proposed a two-dimensional two-moment normalization method [26], which divides the stroke features into eight feature extraction directions. The features in each direction are Gaussian fuzzy processed to ensure a balanced distribution of Japanese character features. The method has the following mathematical equation expression.

$$g_{t}(x, w) = \sum_{j=1}^{k} \frac{1}{\lambda_{ij}} \left[ \varphi_{ij}^{T}(x - \mu_{j}) \right]^{2} + \sum_{j=1}^{k} \log \lambda_{ij} + (n - k) \log \delta$$  (1)

where $\mu_{i}$ denotes the mean vector in the $w_{i}$ direction, $\varphi_{ij}$ denotes the eigenvector of the corresponding covariance matrix $\lambda_{ij}$, $X$ denotes the number of filters, $k$ denotes the constant eigenvector, $n$ denotes the character eigenconstant parameter, and $\delta$ denotes the variable that can be optimized in the training parameters. For Japanese handwriting, we use a decoder mechanism to decompose the hiragana handwriting trajectory. We set the feature decoding handwriting time step as $t$ and the attention weight as $\rho_{t}$. Japanese handwriting and trajectories vary from person to person, so to distinguish scribbles from regular handwriting, we set the implicit computation of the target feature in the encoder feature encoding layer as $h_{t}$. The mathematical expression is shown below.

$$h_{t}, O_{t} = \text{LSTM}(h_{t-1}, a_{t}),$$  (2)

where $O_{t}$ denotes the graded output value of the hidden feature layer in time step $t$ and $a_{t}$ denotes the attentional feature vector in time step $t$. LSTM denotes the transition network between the two hidden layers. Based on the previously trained model, we adopted the attention mechanism parameters of the pretrained model, encoded the trajectory vectors before and after the Japanese characters, and redesigned the mapping relationship of the hiragana labels of the trajectory features according to the feature pointing in the hidden layer. At the tail end of the encoder, we add a softmax layer and finally generate a predictive distribution for Japanese handwritten fonts through the joint action of the encoder and the softmax activation function, whose mathematical equation expressions are shown below.

$$y_{t} = \text{softmax}(W_{a} \times a_{t}).$$  (3)

For feature extraction of Japanese computer-written and handwritten scripts, we used an attention mechanism model. To effectively accomplish the distributed prediction of attention vectors, we try to keep the integrity of the features during the coding process. For the irregularity problem of handwriting style, we store the vectors of different feature directions independently with the network of long short-term memory units. Considering the specificity of trajectory tracking vectors, we use a two-layer memory cell structure to store trajectory information and direction information separately. To prevent repeated prediction of character features, we reconstruct the long short-term memory network by setting a fixed storage length for each memory cell. The softmax activation function generates a predictive distribution of attention vectors for fixed-length memory cells, and the corresponding hiragana features can be matched according to the directional orientation of the attention vectors. The process of hiragana feature extraction by the attention mechanism is shown in Figure 2.

3.3. Japanese Voice Sequence Encoder. In the construction of the Japanese voice recognition system, we summarized previous studies and experimentally demonstrated the methods mentioned in the literature. In the literature [27], a recurrent neural network voice recognition method is proposed, in which the long sequence gradient dependence is first performed for the output voice sequences, and then the voice segments are decomposed by feature gradients in this way, to convert them into mapping links with Japanese hiragana and thus accomplish the task of voice recognition. The literature [28], on the other hand, improves on the former, and the authors propose a voice sequence recognition method with long short-term memory units. The voice sequences are segmented by the gating unit, then the memory unit is used to store the voice sequences, and the
voice sequence recognition is accomplished by matching the mapping with hiragana character features. Researchers in the literature [29] improved based on the memory unit network and proposed a voice sequence recognition method with double-layer memory units, which accelerated the processing speed of voice sequences and realized the Internet online dataset variant processing. Combining the experimental results from the above literature, we used a two-way gated memory cell network structure to recognize Japanese voice sequences.

Researchers in the literature [30, 31] have proposed methods to perform encoding on scene voice transformation and feature mapping. We applied the method to a network of gated memory units. The voice sequence is first segmented, and then the segmented voice fragments are assigned to independent gated units, each corresponding to an independent encoder. To capture the temporal information of the voice sequences, we arrange the gating units and hierarchically traverse the voice feature nodes in each row of the network. In the second feature traversal, we set the order of extraction from high-level features to low-level features until the voice sequences are not repeatedly segmented within the gating units. Our voice sequence processing flow is shown in Figure 3.

3.4. Multi-Source Feature Fusion Mapping System. We design a hybrid intelligent teaching system for the Japanese language with multi-source information fusion mapping, as shown in Figure 4. The system is mainly divided into a text Japanese recognition branch and a Japanese voice recognition branch. In the text recognition branch, we use an attention mechanism to decompose the text information. For Japanese computer-written and handwritten styles, we use different text feature extraction methods and finally feature aggregation is performed by long short-term memory networks. For Japanese voice recognition units, we used memory gating units to segment the voice sequences and then assigned the segmented voice fragments to independent gating units, each corresponding to an independent encoder. The voice sequences will then be automatically extracted by the neural network in the gating unit. We add a double-layer voice sequence memory unit in the network layer to speed up the processing speed of voice sequences.
The dual recognition of Japanese text and voice together constitutes a hybrid Japanese intelligent teaching aid system.

4. Experiment

4.1. Datasets. To verify the effectiveness of our designed Japanese hybrid identification system with multi-source information fusion, we chose the public dataset to launch the experiment. The researchers in literature [32] proposed a Japanese character recognition dataset, Kuzushiji, in which most of the Japanese characters were generated by transcription. This dataset was also expanded in later studies, and most of the expanded data were computer-written Japanese characters. Literature [33] proposed a Japanese voice recognition dataset ASR, which contains more than 2000 hours of Japanese voice content, and most of the Japanese scenes are from Japanese dramas and Japanese life scenes on YouTube. The dataset not only prepares the voice content, but also annotates each voice content with hiragana subtitle labels, and this dataset saves data preprocessing costs for voice recognition work. Details of the dataset are shown in Table 1.

In addition to validating our method on a public dataset, we created our Japanese dataset based on the needs of the application. For the Japanese text recognition branch, we produced a small batch text dataset by manually integrating Japanese textbooks. For the Japanese voice recognition branch, we collated some Japanese drama clips and preprocessed them with voice noise reduction, denoising, and audio track separation, respectively. Then, voice segmentation is performed according to duration and subject categories. The segmented voice sequences are processed by us according to the feature alignment of voice sequences in ASR data. The voice sequence data preprocessing process is shown in Figure 5.

4.2. Experimental Results. We select the same type of text recognition algorithm as a comparison. Recurrent neural network (RNN) [34] is one of the most commonly used algorithms in the field of text recognition. Based on RNN, some researchers have improved the neural network structure and proposed the long short-term memory (LSTM) unit network [35]. For text segmentation, the CTPN algorithm [36] is more advantageous, which is optimized based on Faster RCNN, and it retains the excellent image recognition ability of the CNN family. The main workflow of this algorithm consists of text box detection, text box recurrent concatenation, and text refinement. To validate the text recognition accuracy of our method on public datasets, we test it on the Kuzushiji dataset and a homemade dataset.

| Table 1: Japanese text and voice public dataset details. |
|---------------------------------|---|---|---|
|                                | Train | Test | Total |
| Kuzushiji                      | 35961 | 3995 | 39956 |
| ASR                            | 23451 | 2113 | 25564 |

| Table 2: Results of Japanese text detection by different methods. |
|----------------|---|---|---|---|---|---|
|                 | Kuzushiji |                |                | Self-made datasets |                |
|                 | Acc (%) | P     | E     | Acc (%) | P     | E     |
| RNN             | 71      | 3.4   | 0.32  | 69      | 3.5   | 0.21  |
| LSTM            | 79      | 2.8   | 0.23  | 76      | 2.9   | 0.16  |
| CTPN            | 86      | 2.1   | 0.15  | 86      | 2.1   | 0.11  |
| Ours            | 96      | 0.9   | 0.08  | 93      | 1.1   | 0.05  |
We tested accuracy (Acc), number of parameters (P), and error rate (E). The experimental results are shown in Table 2.

From the above experimental results, it is clear that all the algorithms perform lower overall on the public dataset than on the homemade dataset for Japanese text recognition. The overall recognition results on the public dataset Kuzushiji are poor due to the wide coverage of the public dataset and the large number of Japanese characters involved, in addition to the inclusion of a large number of ancient transcribed Japanese characters. The sample size of our homemade dataset is too small, and the dataset production cost is large, so the homemade dataset is deficient in terms of data volume. In terms of accuracy, our method achieves 96% on the public dataset, which is better than other methods. At the parametric number level, our method has only 0.9. Since our method additionally adds a lightweight structure, the number of parameters is smaller. At the error rate level, the error rate of our method is only 0.08, while the error rates of other methods are all greater than 1, thus proving the effectiveness of our method. In the Japanese text feature set scatter (S) test, we will use the above three algorithms as a comparison. In addition, we also verify the precision (P) and recall (R) of Japanese text recognition, and the experimental results are shown in Table 3.

From the experimental results in the above table, we can see that RNN and LSTM perform poorly in the Japanese text feature scatter test, CTPN achieves 0.8 in the scatter test, and our method achieves 0.9 in the scatter test. According to the Japanese text test accuracy level, our method achieves 96% Japanese text detection accuracy and 97% recall, which is better than other algorithms, proving the effectiveness of our method.

For the Japanese voice recognition branch, we set up a separate experimental verification session. Deep neural networks (DNNs) [37] are more widely used in voice recognition. Based on DNNs, some researchers fused hidden Markov models and proposed a DNN-HMM method for voice sequence recognition [38]. This method can effectively handle long sequence voice sequences compared to DNN methods. Other researchers have proposed the TDNN [39] voice sequence recognition model. The method first performs Fourier transform on the voice sequence and then converts the voice sequence into a signal image, and the output unit directly matches the character results. To validate the accuracy of our method for Japanese voice sequence recognition on public datasets, we test it on ASR datasets and homemade datasets. Our testing criteria are accuracy (Acc), F1 score, and voice sequence segmentation rate. The experimental results are shown in Table 4.

From the experimental results in the above table, it can be seen that our method achieves 93% accuracy on the Japanese voice sequence public dataset, due to all other methods. The F1 score also reaches 0.91. This shows the high efficiency of our method. In the production of the homemade dataset, we used the same production process as the ASR dataset, so the experimental results of the homemade dataset and the ASR dataset are not very different. For the Japanese voice sequence recognition efficiency test, we also added the set dispersion test (S), precision test (P), and recall test (R). The experimental results are shown in Table 5.

From the experimental results in the above table, we can see that our method achieves 0.9 in the set dispersion, and the precision and recall remain above 90%, which is better than other algorithms. All previous experimental results fully demonstrate that our proposed hybrid Japanese intelligent recognition system can achieve good accuracy criteria.

5. Conclusion

We propose a hybrid Japanese teaching aid system with multiple information fusion mapping that can effectively improve the efficiency of Japanese language teaching and reduce the cumbersome human teaching procedures. The system is divided into two branches of Japanese language recognition, the Japanese text recognition branch, and the Japanese voice sequence recognition branch. In the Japanese text recognition branch, we integrate attention mechanisms and long short-term memory networks as the basic network for Japanese character text recognition. In addition, we set up separate text feature recognition systems for Japanese computer-written and handwritten characters to prevent the problem of feature overlap. For Japanese voice sequence recognition, we use a combination of memory gating unit and encoder, based on the network still extending the structure of the deep neural network, and using residual structure connection in the gating unit to avoid the gradient disappearance problem. To verify the efficiency of our system, we selected the Japanese text recognition public dataset and voice recognition public dataset for

| Method     | S  | P (%) | R (%) |
|------------|----|-------|-------|
| RNN        | 0.5| 75    | 79    |
| LSTM       | 0.6| 88    | 86    |
| CTPN       | 0.8| 90    | 90    |
| Ours       | 0.9| 96    | 97    |

| Method     | S  | P (%) | R (%) |
|------------|----|-------|-------|
| DNN        | 0.7| 81    | 81    |
| DNN-HMM    | 0.8| 85    | 86    |
| TDNN       | 0.8| 89    | 90    |
| Ours       | 0.9| 93    | 95    |

Table 3: Japanese text detection experiment results.

| Method     | Acc | F1 | Acc | F1 |
|------------|-----|----|-----|----|
| DNN        | 76% | 0.84| 81% | 0.82|
| DNN-HMM    | 81% | 0.86| 83% | 0.86|
| TDNN       | 89% | 0.89| 91% | 0.89|
| Ours       | 93% | 0.91| 95% | 0.91|

Table 4: Results of Japanese voice sequence recognition by different methods.

| Method     | S  | P (%) | R (%) |
|------------|----|-------|-------|
| DNN        | 0.7| 81    | 81    |
| DNN-HMM    | 0.8| 85    | 86    |
| TDNN       | 0.8| 89    | 90    |
| Ours       | 0.9| 93    | 95    |

Table 5: Japanese voice sequence recognition experiment results.
experimental validation. To match the practical application of the system, we make our dataset according to the dataset standard and perform experimental validation. The experimental results show that our method is significantly better than other methods, and the accuracy and precision are maintained above 90%.

Our hybrid Japanese-assisted teaching system has a special application scenario, so we make our own relevant application Japanese scenario dataset to train the model. Based on the previous experimental results, we can see that our dataset is not perfect and the recognition performance is not accurate enough. In future research, we will further expand the number of homemade datasets and further improve the structural optimization work of the network.

Data Availability

The dataset can be accessed upon request.

Conflicts of Interest

The authors declare that there are no conflicts of interest.

References

[1] M. Shabudin, N. Hieda, and N. Amzah, “Bahasa Jepun sebagai bahasa asing di Universiti Kebangsaan Malaysia. Dlm Ambigapathy Pandian et al.(pnyt.),” Panca Dimensi Pengajaran dan Pembelajaran Bahasa: Trend dan Amalan, vol. 12, pp. 117–127, 2010.

[2] B. Matthew, “Closed-and open-ended narratives of personal experience: weekly meetings among a supervisor and teaching assistants of a “Japanese language education practicum,” Linguistics and Education, vol. 15, no. 1-2, pp. 3–32, 2004.

[3] V. Peltokorpi and L. E. Zhang, “Host country culture and language identification, and their workplace manifestations: a study on corporate expatriates in China and Japan,” Journal of International Management, vol. 28, no. 3, Article ID 100926, 2022.

[4] N. T. Ly, C. T. Nguyen, and M. Nakagawa, “An attention-based row-column encoder-decoder model for text recognition in Japanese historical documents,” Pattern Recognition Letters, vol. 136, pp. 134–141, 2020.

[5] H. Higuchi, S. Iwaki, and A. Uno, “Altered visual character and object recognition in Japanese-speaking adolescents with developmental dyslexia,” Neuroscience Letters, vol. 723, Article ID 134841, 2020.

[6] S. Eskenazi, P. Gomez-Krämer, and J. M. Ogier, “A comprehensive survey of mostly textual document segmentation algorithms since 2008,” Pattern Recognition, vol. 64, pp. 1–14, 2017.

[7] H. Bunke, “Recognition of cursive Roman handwriting: past, present and future,” in Proceedings of the Seventh International Conference on Document Analysis and Recognition, pp. 448–459, Edinburgh, UK, September 2003.

[8] H. Fujisawa, “Forty years of research in character and document recognition—an industrial perspective,” Pattern Recognition, vol. 41, no. 8, pp. 2435–2446, 2008.

[9] Q. F. Wang, F. Yin, and C. L. Liu, “Handwritten Chinese text recognition by integrating multiple contexts[,]” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 34, no. 8, pp. 1469–1481, 2011.

[10] R. Plamondon and S. N. Srihari, “Online and off-line handwriting recognition: a comprehensive survey,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 22, no. 1, pp. 63–84, 2000.

[11] A. Elad and R. Kimmel, “On bending invariant signatures for surfaces,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 25, no. 10, pp. 1285–1295, 2003.

[12] W. Zhang, L. Deng, L. Yang et al., “Multilanguage-handwriting self-powered recognition based on triboelectric nanogenerator enabled machine learning,” Nano Energy, vol. 77, Article ID 105174, 2020.

[13] S. Fine, Y. Singer, and N. Tishby, “The hierarchical hidden Markov model: analysis and applications[,]” Machine Learning, vol. 32, no. 1, pp. 41–62, 1998.

[14] H. D. Yang, S. Sclaroff, and S. W. Lee, “Sign language spotting with a threshold model based on conditional random fields,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 31, no. 7, pp. 1264–1277, 2009.

[15] P. Natarajan, R. Prasad, H. Cao et al., “Arabic text recognition using a script-independent methodology: a unified HMM-based approach for machine-printed and handwritten text,” Guide to OCR for Arabic Scripts, pp. 485–505, Springer, London, 2012.

[16] P. Kahle, S. Colutto, G. Hackl, and G. Mühlberger, “Transkribus-a service platform for transcription, recognition and retrieval of historical documents,” in Proceedings of the 2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR), vol. 4, pp. 19–24, IEEE, Kyoto, Japan, November 2017.

[17] N. T. Ly, C. T. Nguyen, and M. Nakagawa, “Training an end-to-end model for offline handwritten Japanese text recognition by generated synthetic patterns,” in Proceedings of the 2018 16th International Conference on Frontiers in Handwriting Recognition (ICFHR), pp. 74–79, IEEE, Niagara Falls, NY, USA, August 2018.

[18] E. J. Keogh and M. J. Pazzani, “Derivative dynamic time warping,” in Proceedings of the 2001 SIAM international conference on data mining, pp. 1–11, Society for Industrial and Applied Mathematics, Beijing China, July 2001.

[19] G. Tomasi, F. Van Den Berg, and C. Andersson, “Correlation optimized warping and dynamic time warping as preprocessing methods for chromatographic data,” Journal of Chemometrics, vol. 18, no. 5, pp. 231–241, 2004.

[20] S. M. Simmons, J. K. Caird, and P. Steel, “A meta-analysis of in-vehicle and nomadic voice-recognition system interaction and driving performance,” Accident Analysis & Prevention, vol. 106, pp. 31–43, 2017.

[21] A. Zulfiqar, A. Muhammad, A. M. Martinez-Enriquez, and G. Escalada-Imaz, “Text-independent speaker identification using VQ-HMM model based multiple classifier system,” in Proceedings of the Mexican International Conference on Artificial Intelligence, pp. 116–125, Springer, Berlin, Heidelberg, July 2010.

[22] Y. Lin and X. Song, “Order selection for regression-based hidden Markov model,” Journal of Multivariate Analysis, vol. 192, Article ID 105061, 2022.

[23] B. Zhu and M. Nakagawa, “A MRF model with parameter optimization by CRF for on-line recognition of handwritten Japanese characters[,]” in Proceedings of the 2017 International Conference on Document Analysis and Recognition (ICDAR), vol. 4, pp. 19–24, IEEE, Kyoto, Japan, November 2017.

[24] B. Zhu and M. Nakagawa, “A MRF model with parameter optimization by CRF,” in Proceedings of the 2011 International Conference on Document Analysis and Recognition (ICDAR), vol. 4, pp. 19–24, IEEE, Kyoto, Japan, November 2017.
Conference on Document Analysis and Recognition, pp. 603–607, IEEE, Beijing, China, November 2011.

[25] F. Kimura, K. Takashina, S. Tsuruoka, and Y Miyake, “Modified quadratic discriminant functions and the application to Chinese character recognition,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 1, pp. 149–153, 1987.

[26] C. L. Liu and K. Marukawa, “Pseudo two-dimensional shape normalization methods for handwritten Chinese character recognition,” Pattern Recognition, vol. 38, no. 12, pp. 2242–2255, 2005.

[27] S. Hochreiter and J. Schmidhuber, “Long short-term memory,” Neural Computation, vol. 9, no. 8, pp. 1735–1780, 1997.

[28] K. Cho, B. Van Merriënboer, C. Gulcehre et al., “Learning phrase representations using RNN encoder-decoder for statistical machine translation[1],” 2014, https://arxiv.org/abs/1406.1078.

[29] K. Greff, R. K. Srivastava, J. Koutník, B. R. Steunebrink, and J. Schmidhuber, “LSTM: a search space odyssey,” IEEE Transactions on Neural Networks and Learning Systems, vol. 28, no. 10, pp. 2222–2232, 2017.

[30] Y. Deng, A. Kanervisto, J. Ling, and A. M. Rush, “Image-to-markup generation with coarse-to-fine attention,” in Proceedings of the International Conference on Machine Learning, pp. 980–989, PMLR, NY City, June 2017.

[31] R. Geetha, T. Thilagam, and T. Padmavathy, “Effective offline handwritten text recognition model based on a sequence-to-sequence approach with CNN–RNN networks[1],” Neural Computing & Applications, vol. 33, no. 17, pp. 10923–10934, 2021.

[32] Z. Tian, W. Huang, T. He, P. He, and Y. Qiao, “Detecting text in natural image with connectionist text proposal network,” in Proceedings of the European conference on computer vision, pp. 56–72, Springer, Amsterdam, The Netherlands, October 2016.

[33] K. Aizat, O. Mohamed, M. Orken, A. Ainur, and B. Zhumazhanov, “Identification and authentication of user voice using DNN features and i-vector,” Cogent Engineering, vol. 7, no. 1, Article ID 1751557, 2020.

[34] J. Li, D. Yu, J. T. Huang, and Y. Gong, “Improving wideband speech recognition using mixed-bandwidth training data in CD-DNN-HMM,” in Proceedings of the 2012 IEEE Spoken Language Technology Workshop (SLT), pp. 131–136, IEEE, Miami, FL, USA, December 2012.

[35] R. V. Sharan, U. R. Abeyratne, V. R. Swarnkar, and P. Porter, “Automatic croup diagnosis using cough sound recognition,” IEEE Transactions on Biomedical Engineering, vol. 66, no. 2, pp. 485–495, 2019.