Journal of Artificial Intelligence Practice (2023)
Clausius Scientific Press, Canada

DOI: 10.23977/jaip.2023.060506
ISSN 2371-8412 Vol. 6 Num. 5

Optimization and Evaluation of Spoken English CAF Based on Artificial Intelligence and Corpus

Wenfang Zhang[1,a,*], Xiaodong Wang[2,b]

[1]School of Language and Culture, Graduate University of Mongolia, Ulaanbaatar, Mongolia
[2]School of Government Management, Inner Mongolia Normal University, Hohhot, Inner Mongolia Autonomous Region, China
[a]1092840896@qq.com, [b]339626878@qq.com
[*]Corresponding author

Keywords: Spoken English CAF, Artificial Intelligence, HMM Model, CNN Model

Abstract: English is the most widely used language in the world, and the pronunciation of its spoken language is equally important. The traditional methods are not high in complexity, accuracy and fluency (CAF) for spoken English recognition. Therefore, it is very important to use AI and corpus to optimize and evaluate spoken English CAF. This paper aims to study the optimization and evaluation of spoken English CAF using AI and corpus, and proposes to use the Hidden Markov (HMM) model and convolutional neural network (CNN) model in the field of AI to optimize and evaluate spoken English CAF. By selecting a variety of English voices from the BNC corpus for model training and testing, and selecting the complexity, accuracy, fluency and harmonic average of the CNN model recognition as evaluation indicators, the HMM model's recognition spectrogram is added up and analyzed. In the experimental test, it was found that when the number of frames is 210, the indicators of the CNN model have been greatly improved, so the number of frames selected for the test in this paper is 210. The results show that the A value obtained by the HMM model test is about 85%, the CNN model is 67%, and the traditional SVM model is only 35%. The HMM model is tested with a C value of about 60%, the CNN model is 65%, and the traditional model is only 45%. The F-value obtained from the test of the HMM model is about 83%, the CNN model is 67%, and the traditional model is 46%. In contrast, the HMM model has higher recognition accuracy for spoken English, and the recognition results are more fluent. However, the CNN model can recognize spoken English with higher complexity, and both the CNN model and the HMM model can improve the CAF optimization effect of spoken English.

1. Introduction

English is the most widely used language in the world. In the era of global economic integration, countries have higher and higher requirements for spoken English CAF. It is more difficult to recognize spoken English for non-specific people. To truly realize non-specific spoken English recognition, a large amount of data collection needs to be carried out in the data collection stage, including sufficient voice collection for people of all ages and English levels. However, the

traditional machine learning method is not very effective in identifying CAF in spoken English, and it is more difficult to meet the requirements in complex environments such as non-specific people. Therefore, more intelligent and efficient CAF optimization methods are required.

Some scholars have put forward different ideas and methods in order to optimize and evaluate the CAF of spoken English. Among them, Kim N studied and investigated the effects of unplanned and online planning on the complexity, accuracy and fluency (CAF) of L2 spoken and written tasks [1]. Economidou-Kogetsidis M tracked the longitudinal pragmatic development of English learners' oral requests for an academic year abroad, which aimed to examine whether and how their request performance developed over time in high- and low-intensity situations [2]. Liu CY used the Academic Spoken Word List (ASWL) and the British National Corpus and Corpus of Contemporary American English (BNC/COCA) lists to analyze the lexical profiles of TED talks [3]. Stange U explored the discursive use of selected emotional interjections in spoken British English. [4]. Sudar S studied the use of discourse patterns by English teachers and their students in the classroom. And a qualitative descriptive study was used to analyze the data [5]. However, these methods have low efficiency, insufficient recognition accuracy and small application range.

In order to solve the problem of low efficiency, some scholars have proposed methods by using artificial intelligence and corpus. Among them, English reading plays an important role in improving oral English and comprehensive English ability. Meng Q built a system that based on artificial intelligence algorithm and combined with spoken language spectrum algorithm [6]. Baniulyte G conducted electronic searches and research of spoken English corpora on PubMed, Cochrane, Scopu and other data sources. All studies showed an overall positive trend in AI technology [7]. Ozn G reported the construction of a 240,000-word English (CPE) pilot corpus. The categories of articles and the proportions of monologues and dialogues were guided by the International Corpus of English Project, which made the corpus directly comparable to existing postcolonial English corpora [8]. Although the methods of these studies improve efficiency, they are costly and difficult to implement.

This paper adopts the HMM model and CNN model to optimize and evaluate the CAF of spoken English. The HMM model uses a maximum probability matching algorithm to identify spoken English, while the CNN model uses feature extraction to identify spoken English. The accuracy rate alone cannot judge the model performance, so this paper compares and analyzes the complexity (C), accuracy (A), fluency (F) and harmonic mean (AVG) indicators. The test found that the category with a large number of samples also achieved better results in the test, and it also proved that the sample imbalance problem did bring a negative impact on the experiment when training the model. The frame number value of 210 is applied to the English accent recognition experiment of the traditional SVM model, and the accuracy of the traditional model is between 28% and 42%, and the average accuracy is 35%. The accuracy rate of the CNN English accent recognition model is between 60% and 74%, and the average accuracy rate is 67%, which is much higher than the SVM model. In addition, it is obtained from the HMM model to identify the spectrum, the HMM identification result is not much different from the actual frequency map, and it can be said that it is basically the same. The main error is still in the high frequency part, which may be because there are more noise points in the high frequency part, which leads to an increase in the recognition error. The overall results show that both the CNN model and the HMM model can achieve good CAF optimization results for spoken English.

With the development of AI technology, the previous methods can no longer meet the current environment. The innovation of this paper is that the CNN neural network model and HMM model in the AI field are proposed to optimize the spoken English CAF, and the speech information in the BNC corpus is selected as the source of data.

2. Oral CAF Optimization Method Based on AI and Corpus

2.1 Spoken English CAF

Spoken English CAF is an abbreviation of the complexity, accuracy and fluency of oral English expression, and it has been widely concerned by scholars at home and abroad in the research of English expression and conception. As early as 2013, researchers proposed an index system to measure language complexity, accuracy and fluency. In order to measure the development of the CAF triplet of the learner's language, some researchers have adopted a new cognitive strategy, that is, the average length of discourse flow is an important measurement index. In fact, such measurement methods themselves have several shortcomings. Moreover, it does not fully conform to the English spoken language measurement index system. With the continuous deepening of language comparison and interlanguage research, it is more urgent to construct intelligent measurement indicators in the field of spoken English [9-10].

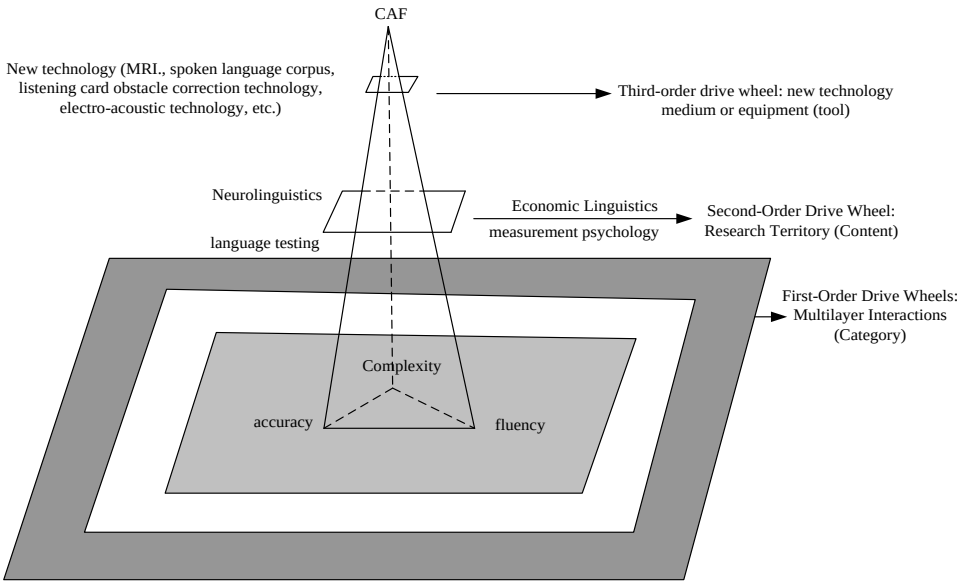


Figure 1: Basic Optimization Ideas of Spoken English CAF

The basic optimization method for spoken English CAF is roughly shown in Figure 1. It is divided into category layer, content layer, tool layer and CAF target layer. The category layer includes the pronunciation rules of spoken English in different regions and locales. The content layer is the comprehensive research and application of various disciplines. The tool layer assists the testing of various new technology media and software devices. The target layer is the final CAF optimization for spoken English [11].

The three dimensions of CAF were basically used to measure learners' spoken English and spoken English at first. So far, in the field of English acquisition, a large number of related studies have taken language complexity, accuracy and fluency (CAF) as a dependent variable, and measured learners' spoken English proficiency by means of three-dimensional measures of spoken English. Language complexity refers to a higher level of linguistic thought processing. The accuracy of language can be understood as the language habit that is free from grammatical errors and conforms to the target language. Fluency is equated with the speed of language production throughout the oral English conceptualization process. However, in the process of using spoken language, the understanding of the CAF triplet is not so simple. Generally speaking, the CAF triplet itself is a multi-dimensional dynamic real-time monitoring process, which includes more complex

conceptual connotations and quantitative systems. Even domestic and foreign English acquisition researchers have not completely consistent with their conceptual definitions and quantification systems. Accordingly, researchers use different index systems to measure each dimension. Among them, the complexity of language is the most difficult to define. It not only refers to the internal ontology representation presented by the English oral task, but also refers to the measurement dimension of the coefficient index in the process of English production. In addition, the complexity of language aims to characterize language output. The conceptual connotation of the dimension of complexity still has multiple meanings, because it can be used in multiple interfaces of language production and language communication [12-13].

There are generally three meaning orientations of complexity: the first is the syntactic complexity in the language text; the second is the diversity of the sentence structure of the language text; the third is based on the process order of language acquisition, that is, languages acquired later are usually more complex and obscure. In addition, linguistic accuracy is the simplest and most internally consistent construct among the three dimensions, which mainly refers to "the degree of conformity with a specific standard" or "the ability to produce error-free language". In second language acquisition, fluency is often understood by generalization, which is equivalent to language proficiency under certain contextual constraints. For example, if someone speaks English fluently, that person is generally considered to have a high level of English overall. As an important dimension of language production performance, fluency is defined as "the ability to produce language at a normal pace without interruption", or "the validity of language production in the time remaining except for necessary pauses". Both of these definitions are characterized as a certain reference standard. Generally speaking, they mainly use the pause behavior of native speakers as an important reference standard [14-15].

2.2 Artificial Intelligence and Corpus

AI is a branch of computer science, including a variety of algorithms and mathematical models in the field of AI, among which CNN neural network is one of them. The AI oral English CAF evaluation method adopted in this paper, its theoretical idea is shown in Figure 2. It is mainly divided into the establishment of evaluation indicators, the calculation of evaluation models and the optimization of CAF [16].

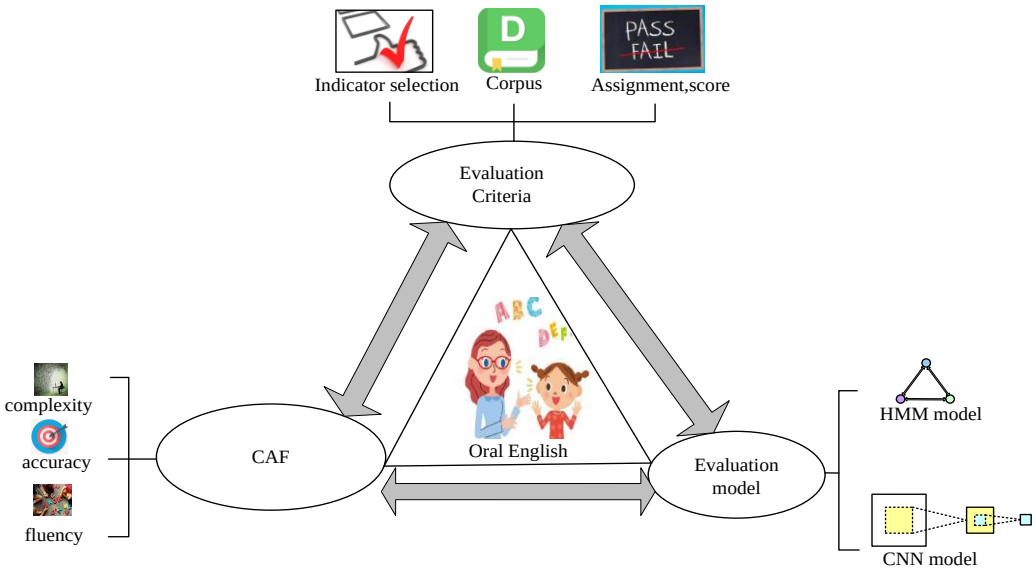


Figure 2: Theoretical Framework of AI Oral English CAF Assessment

CNN has the function of feature extraction. For spoken English, it is the most widely used tool in human communication. Due to the influence of gender, age, region and other factors, there are obvious differences in the pronunciation of different people. These differences in pronunciation provide the possibility for feature extraction and classification of speech data. When human beings observe things, especially when judging the category of things, they often do not need to observe all the characteristics of things before making a conclusion. Rather, the category of things can be determined by observing some of their local manifestations. For example, the judgment is only needed to observe the back of the animal to distinguish whether an animal on a picture is a horse or a camel. The words with insulting content in the text are needed to observe to determine whether a piece of text has an insulting connotation. For different English accents, by comparing American and British accents, the most obvious difference is that there are many reflexes in American accents, but there is no such phenomenon in British accents. Therefore, when distinguishing the American pronunciation of an English pronunciation from the British pronunciation, the characteristic of the reflex can be used as an indicator of the distinction. After capturing the features in speech that characterize these unique pronunciations, these features can be used as the basis for distinguishing different accents. And these features often only exist on some small local segments in speech [17-18].

The design of CNN is inspired by the way the human nervous system transmits and processes information. It is a network structure model that is composed of multiple neurons stacked And by changing the weights of the interconnected neurons, the network structure model of the feature information relationship is obtained. The structural model of the neural network is shown in Figure 3, which is mainly composed of input and output and neurons [19].

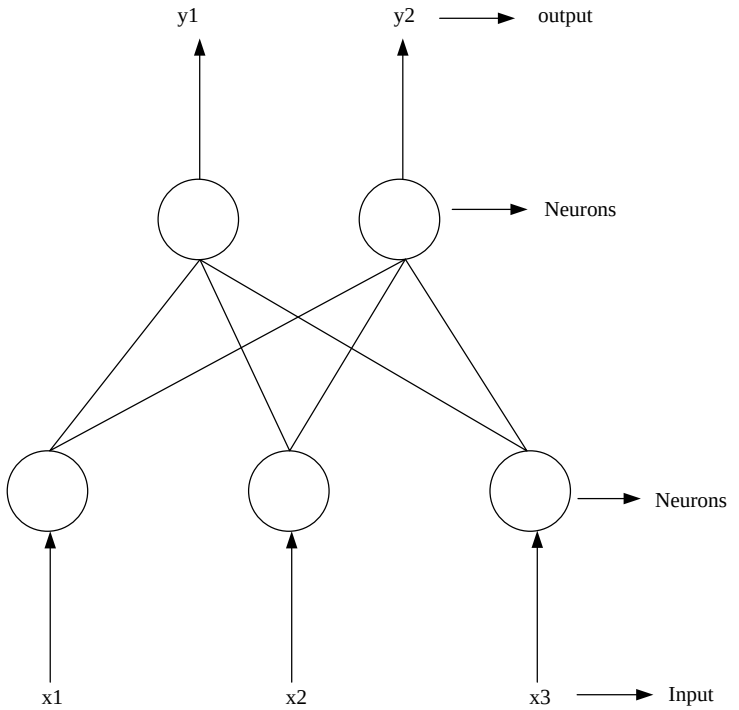


Figure 3: Schematic Diagram of CNN Structure

The recognition architecture of the spoken English CNN model is shown in Figure 4. It can be divided into a training phase and a testing phase. In the training phase, speech is selected from the English corpus for training, and the test is performed after training to obtain a certain accuracy. During the test, the testers read the English material on the spot, and then the speech was obtained

through the model for recognition. The specific calculation process of CNN is not be described in detail in this article, because the CNN model is widely used, and the calculation method is simple and easy to understand and can be seen [20].

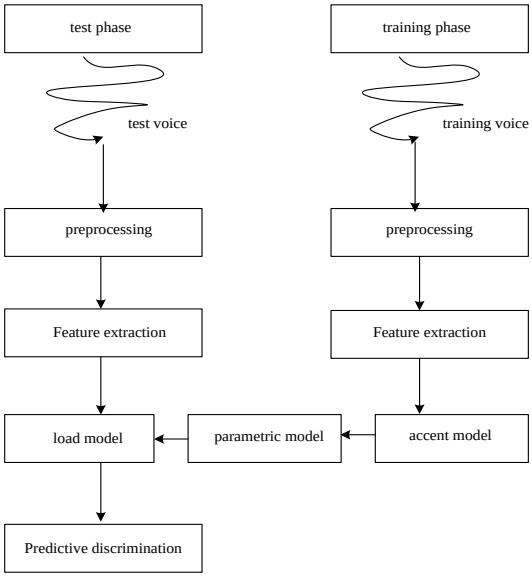


Figure 4: Spoken English CNN Model Recognition Architecture

The speech recognition process of CNN is the recognition of one-dimensional sequence data from the frame level. From the feature level, each frame of speech data contains a series of acoustic features. Therefore, from the perspective of the whole speech feature level, a speech sample is extracted as a two-dimensional feature. One-dimensional convolution of a frame of speech features can only extract the information of the features within the frame, and the most important feature of the speech data is between the frame and the frame data, so the one-dimensional convolution cannot meet the requirements of speech feature extraction.

Two-dimensional convolution is widely used in computer vision and image processing. An example of two-dimensional convolution is shown in Figure 5. It performs a convolution operation according to the input cell size to obtain feature cells and record and use them for reconvolution or pooling. CNN can extract local information in the data based on its convolution operation in sequence feature relation extraction. This sequence relationship extraction method is in line with the local feature extraction of special pronunciations in English accents [21].

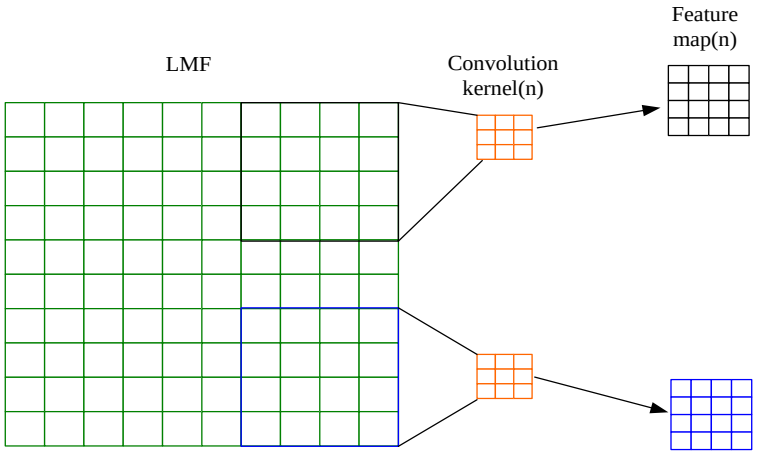


Figure 5: CNN-based English Speech Feature Extraction

It can be seen that CNN extracts the features within the coverage of the convolution kernel in the data through its unique convolution operation. These features are often local features, and then these features are combined at a high level to form features with certain sequence attributes. Therefore, it is possible to identify or evaluate spoken English.

In addition, in the process of constructing the corpus, the feature parameters should be extracted from several speeches read by different people first. Then vectors are quantized to produce a series of characteristic symbols for the speech, which are used for HMM training. Once the speech database and features are obtained, it is possible to start developing a library of speech templates and dictating the speech templates by using the speech in the monophonic database. The training process needs to continuously adjust the parameters of the system template to make the speech recognition system reach an ideal state. In this way, the performance of the system can always be close to this optimal state. In order to save time, this paper directly selects the BNC English corpus as the data source of the experimental test.

2.3 HMM Model

The HMM model is a commonly used speech recognition model, and Figure 6 describes the overall structure of the HMM model. As far as the HMM speech recognition calculation is concerned, the goal is to find the best correlation between the feature parameters x of the speech sequence and the word sequence w generated by the recognition, so that w has a large posterior probability to x.

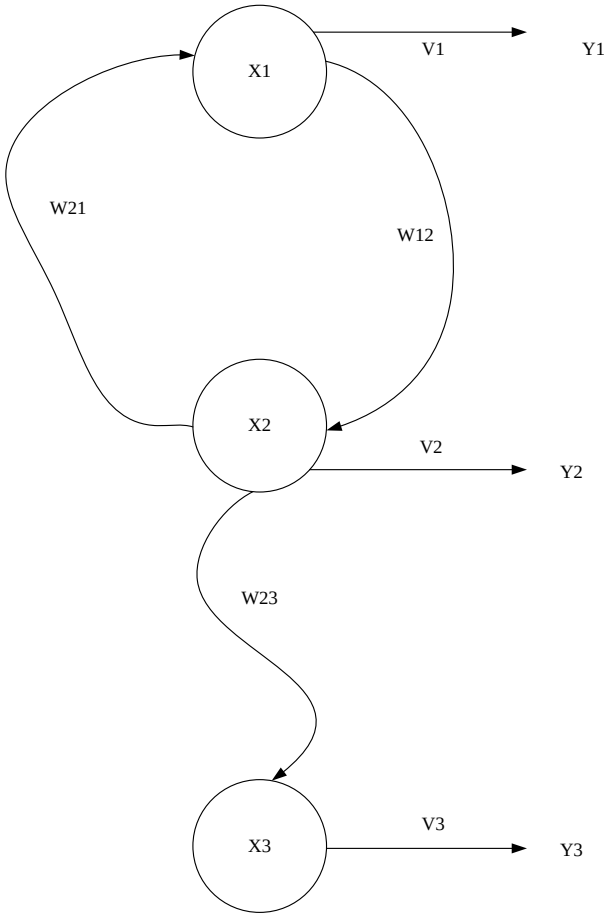


Figure 6: Schematic Diagram of the HMM Model Structure

By using this model, the optimal word sequence w is expressed as:

$$w = \arg\max_{w} \frac{P(x|w)P(w)}{P(x)} \tag{1}$$

Among them, $P(x)$ is only related to the speech input signal and is constant for any recognition sequence, so it can be abbreviated as:

$$w = \arg\max_{w} P(x|w)P(w) \tag{2}$$

It is assumed that the output sequence is a model sequence in the form of an HMM, all alternative paths must be computed and compared. Because each model has several states, each model sequence must account for all conceivable sequences of states that start and end at different times. Therefore, the ideal model sequence can be calculated as the sum of the probabilities of each state sequence occurring. The Formula is expressed as:

$$P(x|w) = \sum_{s} P(x, S|w) \tag{3}$$

In actual calculation, it is very unrealistic to count the probability scores of all state sequences, which means huge computational overhead. In order to reduce the amount of calculation, it can be assumed that the output probability of the HMM model for a piece of speech is represented by the probability value of an optimal state path as:

$$P(x|w) = \max_{S} P(x, S|w) \tag{4}$$

HMM calculation process: according to the KNN criterion, S is divided into J subsets. When $X \in S_i^m$, there are:

$$d(X, Y_I^{m-1}) \le d(X, Y_i^{m-1}) \forall i \tag{5}$$

D represents the deviation, then the total deviation is:

$$D^m = \sum_{i}^{J} \sum_{x \in S_i^m} d(X, Y_i^{m-1}) \tag{6}$$

The relative value of the bias improvement δ^m is:

$$\delta^m = \frac{\Delta D^m}{D^m} = \frac{D^{m-1} - D^m}{D^m} \tag{7}$$

The new codeword is calculated:

$$\left\{ Y_1^m, Y_2^m, ..., Y_J^m : Y_I^m \right\} = 1/N_1 \sum_{x \in S_i^m} X \tag{8}$$

Finally δ^m and m are judged. When $m \ge L$, it stalls, the iteration ends, and $l(l \in (1.J))$ is output.

The training method of the HMM model: the recognition function is represented by g, and the learning sample is represented by X, then it can be defined as:

$$g(X_{k,n}, \vec{\theta}) = \log\left\{ \sum_{all, S_t} P(X_{k,n}, S_c | \vec{\theta}_c)^{\xi} \right\}^{\frac{1}{\xi}} \tag{9}$$

Among them, S_c represents the c-type state sequence, ξ is a normal number, and K represents the sample category.

A simple form recognition function is defined with the Viterbi score as follows, where S_c^* represents the c-class Viterbi best state sequence.

$$g(X_{k,n},\vec{\theta}) = \log(P)((X_{k,n},S_c^*\left|\vec{\theta}_c\right.) \tag{10}$$

The training error of the HMM model can be calculated as:

$$d(X_{k,n},\vec{\Theta}) = -g(X_{k,n},\vec{\theta}) + \left[\frac{1}{M}\sum_p g(X_{k,n},\vec{\theta}_P)^\eta\right]^{\frac{1}{\eta}} \tag{11}$$

In the Formula, M is the number of misidentifications, and η is the weighting coefficient of the misidentification class identification function, which is a positive number.

If $\eta \to \infty$, the error is:

$$d(X_{k,n},\vec{\Theta}) = -g(X_{k,n},\vec{\theta}_k) + g(X_{k,n},\vec{\theta}_c) \tag{12}$$

The loss function of the HMM model can be calculated as:

$$l(d(X_{k,n},\vec{\Theta})) = \frac{1}{1+\exp((-\alpha d)((X_{k,n},\vec{\Theta}))} \tag{13}$$

The total loss is obtained as:

$$L = (X,\vec{\Theta}) = \sum_{k=1}^{K}\sum_{n=1}^{N_n} l(d(X_{k,n},\vec{\Theta})) \tag{14}$$

The calculation of learning iteration can be expressed by Formula (15):

$$\vec{\Theta}(n+1) = \vec{\Theta}(n) - \varepsilon_{\vec{\Theta}(n)}\nabla L((X,\vec{\Theta}(n)) \tag{15}$$

The iterative update method is:

$$\Delta\mu(n+1) = m\Delta\mu(n) + \varepsilon_\mu(n)\frac{\partial L(\mu(n))}{\partial\mu(n)} \tag{16}$$

Among them, m is the parameter for adjusting the inertia rate, which is a constant. ε_μ is the learning coefficient of the mean vector, and the coefficient is:

$$\varepsilon_\mu(n) = \frac{c}{a+bn}\max\frac{\|\mu\|}{\left\|\frac{\partial L(\theta)}{\partial\mu}\right\|} \tag{17}$$

The learning coefficient ε_λ of the eigenvalues is:

$$\varepsilon_\lambda(n) = \frac{c}{a+bn}\max\frac{\|\lambda\|}{\left\|\frac{\partial L(\theta)}{\partial\lambda}\right\|} \tag{18}$$

3. AI Spoken CAF Optimization and Evaluation

3.1 CNN Model Test

As the sequence data, the voice of different durations contains different amounts of information. At the same time, the features of different frame numbers are extracted from the voice data to form feature matrices containing different amounts of information. Too few frames make it impossible to extract features that can truly characterize accents, and too many frames introduce additional noise features. Considering the duration of the experimental sample data and the time-consuming problem of the training process, different characteristic frame numbers F are used to recognize spoken English during the experiment, and the most suitable frame number is found by comparing the results of the experiment.

The corpus used in this paper is the BNC corpus. Because the accuracy rate alone cannot judge the performance of the model, this paper compares and analyzes the experimental results through the indicators of complexity (C), accuracy (A), fluency (F) and harmonic mean (AVG). The experimental test environment is shown in Table 1.

Table 1: CNN Model Experimental Test Environment

Items	Parameter
CPU	Inter Core i9 64bit
Main frequency	3.0GHZ
GPU	NVIDIA GeForce RTX 3060
Programming language	Python

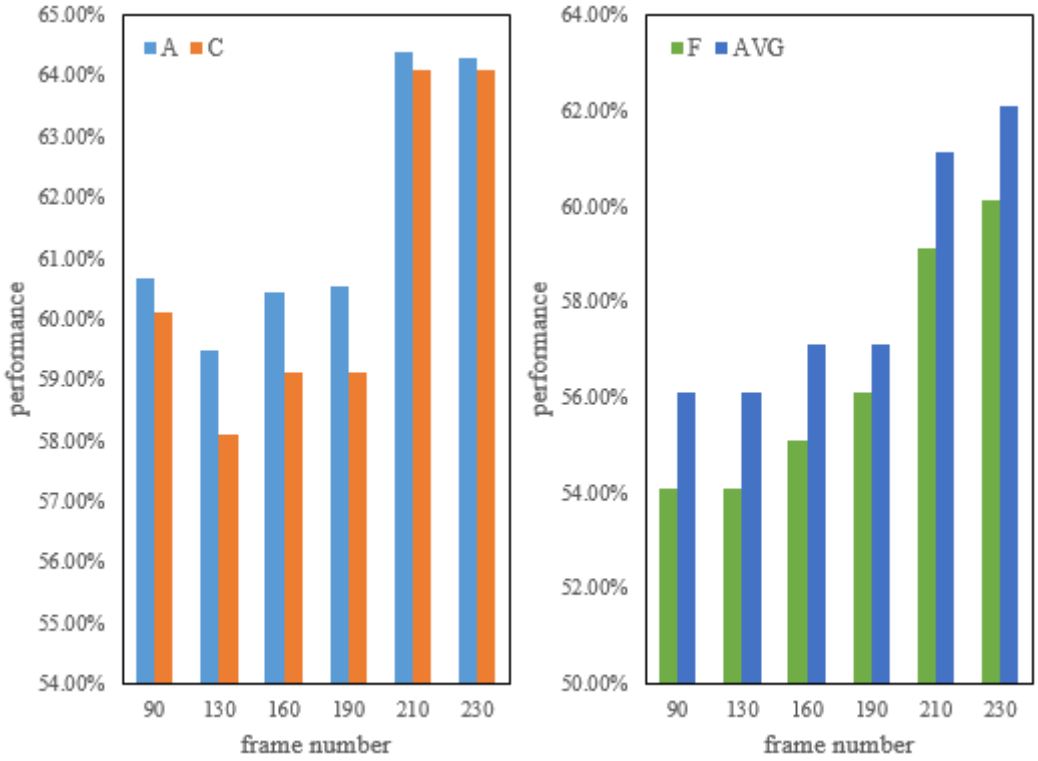


Figure 7: CNN Model Recognition Results for Different Frame Numbers

By comparing the characteristics of different frame numbers in the experiment, the experimental results are shown in Figure 7. It is not difficult to see from Figure 7 that as the number of frames increases, the model uses more and more features, and the model performance also improves. When

the frame number is set to 210, all the indicators of the experimental results have been greatly improved. When the frame number is set to 230, the experimental effect is almost unchanged, especially the ACC, which is slightly reduced. The model structure and the speed of convergence during training are considered. And after the frame number is set to 210, the model performance has almost no change, and the final value of the frame number is set to 210.

Therefore, when the number of frames is 210, this paper summarizes the confusion matrix results of the CNN model on the test set, as shown in Tables 2 and 3. Among them, A, B, C, D, E, F respectively represent the categories of different pronunciations, and x and y represent the horizontal and vertical directions of the matrix.

Table 2: CNN Spoken English Recognition Matrix

reality forecast	Ay	By	Cy	Dy	Ey	Fy	sample number
Ax	1022	28	237	104	212	44	1647
Bx	100	209	58	10	72	9	458
Cx	185	22	1142	71	171	33	1624
Dx	136	13	76	360	76	14	675
Ex	183	16	191	67	1077	30	1564
Fx	76	17	69	27	80	204	473

Table 3: CAF Evaluation Results of CNN Recognition Model

reality forecast	C	A	F	Sample number
A	0.67	0.62	0.6	1647
B	0.65	0.67	0.69	458
C	0.67	0.67	0.67	1624
D	0.66	0.6	0.67	675
E	0.63	0.74	0.68	1564
F	0.64	0.61	0.61	473

From Table 2 and Table 3, it can be seen that the three types of accent recognition accuracy of B, C, and E are relatively high, which are 0.67, 0.67 and 0.74, respectively. It shows that the category with a large number of samples also achieves better results in the test, and it also proves that when the model is trained, the problem of sample imbalance does have a negative impact on the experiment.

Table 4: SVM Spoken English Recognition Data Matrix

reality forecast	Ay	By	Cy	Dy	Ey	Fy	sample number
Ax	826	3	310	19	474	19	1651
Bx	99	209	53	21	62	5	449
Cx	185	14	1138	67	165	39	1608
Dx	137	7	77	360	75	14	670
Ex	173	13	196	73	1086	30	1571
Fx	80	9	70	76	33	82	350

At the same time, in order to compare the effect of traditional machine learning on accent recognition, this paper uses SVM to conduct experiments on the feature matrix of the optimal number of frames. It is different from the two-dimensional convolution of CNN. In the experiment based on SVM, the feature matrix is firstly tiled to form a one-dimensional feature. By considering that the dimension of the one-dimensional feature after tiling is too high, the feature reduction based on principal component analysis is performed on this feature. In the experiment, the feature dimension after dimension reduction is 60, and the obtained results are shown in Table 4 and Figure

8.

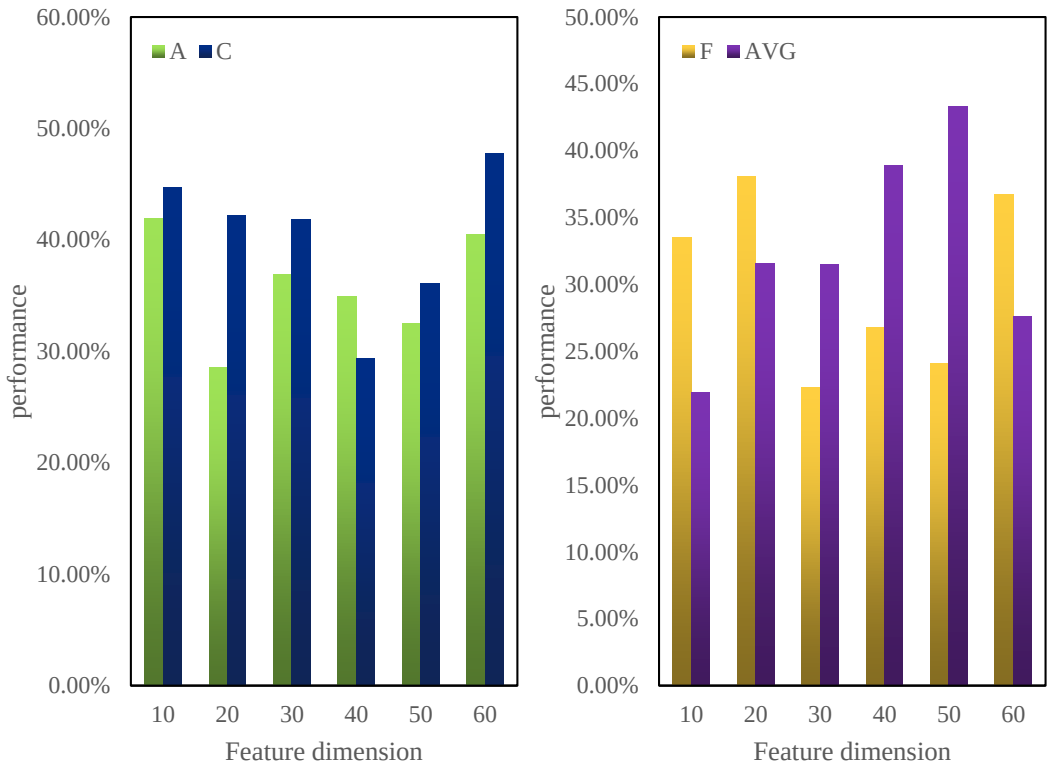
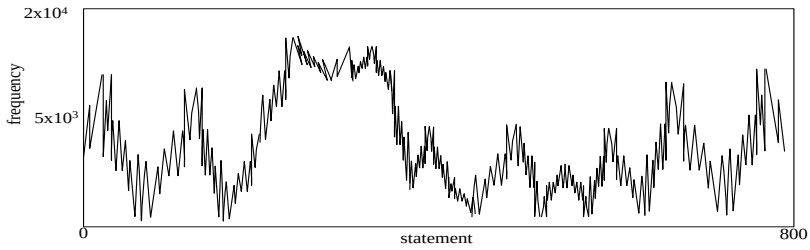


Figure 8: Recognition Results of the Traditional SVM Model

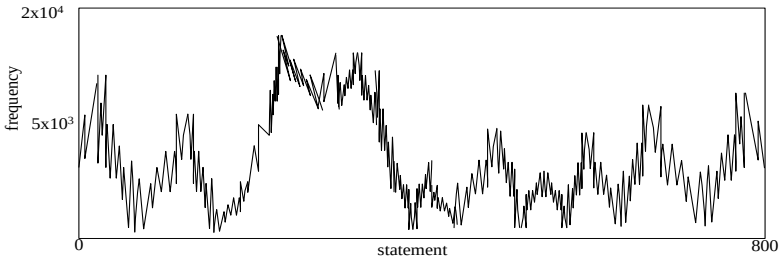
It can be seen from Table 4 and Figure 8 that the frame number value of 210 is applied to the frame number value of the English accent recognition experiment based on SVM. The accuracy of this model is between 28% and 42%, with an average accuracy of 35%. The accuracy rate of the CNN English accent recognition model is between 60% and 74%, and the average accuracy rate is 67%, which is much higher than the SVM model. It shows that the CNN model has a better recognition effect.

3.2 HMM Model Test

In the HMM test, this paper uses words as the basic unit of the model training. The recognition test is performed by selecting 800 speech sentences from the BNC corpus, and the test results are shown in Figure 9.



(a) Actual Speech Spectrogram



(b) HMM Speech Recognition Spectrogram

Figure 9: Speech Spectrogram Comparison

It can be seen from Figure 9 that the spectrogram identified by the HMM is not much different from the actual frequency map, which can be said to be basically the same. The most important error is still in the high frequency part, which may be because there are more noise points in the high frequency part, which leads to an increase in the recognition error. In addition, in order to understand the test results of this model, its accuracy, complexity and fluency are calculated and summarized, and compared with the CNN model and the traditional SVM model, the results are shown in Figure 10.

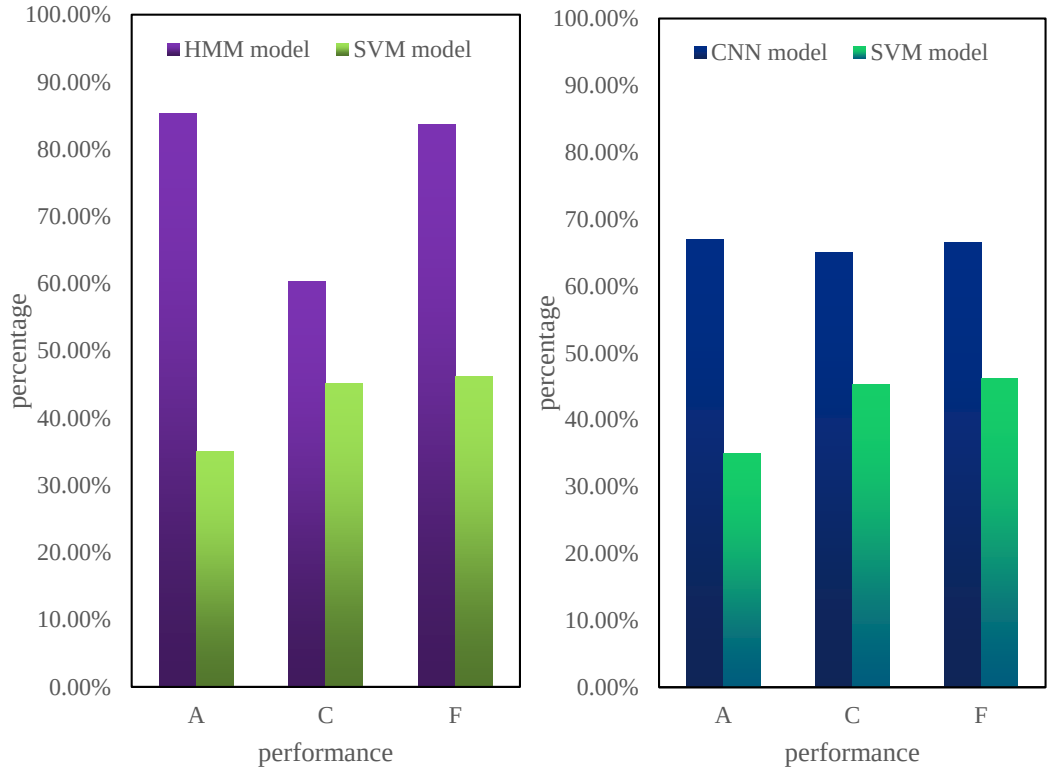


Figure 10: Comparison of Recognition Effects Between HMM, CNN and Traditional Models

It can be seen from Figure 10, the A value obtained by the test of the HMM model is about 85%, the CNN model is 67%, and the traditional model is only 35%. The C value obtained from the test of the HMM model is about 60%, the CNN model is 65%, and the traditional model is only 45%. The F value obtained from the test of the HMM model is about 83%, the CNN model is 67%, and the traditional model is 46%. It can be seen that the HMM model has higher recognition accuracy for the system, and the recognition results are more fluent, and the CNN model can recognize spoken English with higher complexity.

4. Conclusions

This paper firstly summarizes the overall content of the full text in the abstract. Secondly, in the introduction, the development background of AI and the related concepts of CAF in spoken English are introduced, some related researches is cited. In this way, the current situation of the relevant content of this paper can be understood , and the research process and innovation points are summarized. Thirdly, in the theoretical research part, the basic optimization ideas of oral English CAF and the theoretical framework of AI oral English CAF evaluation are introduced. Finally, in the experimental part, the CNN model is first tested, and it is concluded that the frame number value of 210 is applied to the SVM-based English accent recognition experiment, and the accuracy of the model is much higher than that of the SVM model, which indicates that the CNN model has better recognition effect. Among them, the test of the HMM model shows that the A value obtained by the HMM model test is about 85%, the C value is about 60%, and the F value is about 83%. It shows that the HMM model has high recognition accuracy for the system, and the recognition results are relatively fluent. Therefore, the HMM model and CNN model designed in this paper can obtain a good CAF optimization effect for spoken English.

References

[1] Kim N. The Effects of Online Planning on CAF in L2 Spoken and Written Performance. English Teaching, 2018, 73(3):3-28.

[2] Economidou-Kogetsidis M, Halenko N. Developing spoken requests during UK study abroad:A longitudinal look at Japanese learners of English. Study Abroad Research in Second Language Acquisition and International Education, 2022, 7(1):23-53.

[3] Liu CY, HJ Chen. Academic Spoken Vocabulary in TED Talks: Implications for Academic Listening. English Teaching & Learning, 2019, 43(4):353-368.

[4] Stange U. The social life of emotive interjections in spoken British English. Scandinavian Studies in Language, 2019, 10(1):174-193.

[5] Sudar S. Spoken Discourse Analysis of Senior High Schools English Classroom Purworejo, Central Java. Arab World English Journal, 2017, 8(1):194-207.

[6] Meng Q, Tang L. An artificial intelligence based construction and application of English multimodal online reading mode. Journal of Intelligent and Fuzzy Systems, 2020, 40(1):1-10.

[7] Baniulyte G, Ali K. Artificial intelligence - can it be used to outsmart oral cancer? Evidence-Based Dentistry, 2022, 23(1):12-13.

[8] Ozn G, Ayafor M, Green M, Fitzgerald S. The spoken corpus of Cameroon Pidgin English. World Englishes, 2017, 36(3):427-447.

[9] Zhang W, Liu M. Evaluating the Impact of Oral Test Anxiety and Speaking Strategy Use on Oral English Performance. Journal of Asia Tefl, 2017, 10(2):115-148.

[10] Gray S, Restrepo M A, Yeomans-Maldonado G, Bengochea A, Mesa C. The Dimensionality of Oral Language in Kindergarten Spanish–English Dual Language Learners. Journal of Speech Language and Hearing Research, 2018, 61(11):2779-2795.

[11] Meng Q. A Study on Cultivating College Students' Oral English Ability Based on Computer Assisted Language Learning Environment. Boletin Tecnico/technical Bulletin, 2017, 55(4):80-85.

[12] Seraj P, Habil H, Hasan M K. Investigating the Problems of Teaching Oral English Communication Skills in an EFL context at the Tertiary Level. International Journal of Instruction, 2021, 14(2):501-516.

[13] Zeng Y. Application of Flipped Classroom Model Driven by Big Data and Neural Network in Oral English Teaching. Wireless Communications and Mobile Computing, 2021, 2021(1):1-7.

[14] Opeifa O, Adelana O P, Atolagbe O D. Teaching oral English through technology: Perceptions of teachers in Nigerian secondary schools. International Journal of Learning and Teaching, 2022, 14(1):57-68.

[15] Wu H, Sangaiah A K. Oral English Speech Recognition Based on Enhanced Temporal Convolutional Network. Intelligent Automation and Soft Computing, 2021, 28(1):121-132.

[16] Wijewardene L. Oral English Communication Expectations of Business Graduates in the Workplace in Sri Lanka. Advances in Social Sciences Research Journal, 2021, 8(3):104-114.

[17] Tipmontree S, Tasanameelarp A. Using Role Playing Activities to Improve Thai EFL Students' Oral English Communication Skills. International Journal of Business and Society, 2021, 21(3):1215-1225.

[18] Wu H, Ekstam J M. Beyond Parroting: Using English Fun Dubbing to Improve English Oral Performance. Chinese Journal of Applied Linguistics, 2021, 44(2):203-218.
[19] Song Y. The Influence of Background Music Teaching on Accuracy and Fluency of Freshmen's Oral English in China. International Journal for Innovation Education and Research, 2020, 8(11):265-275.
[20] Hai Y. Computer-aided teaching mode of oral English intelligent learning based on speech recognition and network assistance. Journal of Intelligent and Fuzzy Systems, 2020, 39(4):5749-5760.
[21] Lin Y, Ji Q. Analysis of College Oral English Class Design from the Perspective of TBLT—Taking "Read All about It" as an Example. Open Access Library Journal, 2020, 07(11):1-9.