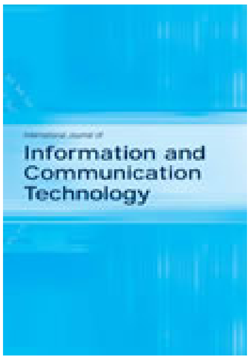


International Journal of Information and Communication Technology

ISSN online: 1741-8070 - ISSN print: 1466-6642
https://www.inderscience.com/ijict

Research on scoring mechanism of spoken English self-study system taking into account artificial intelligence technology and speech knowledge recognition algorithm

Ning Li

DOI: 10.1504/IJICT.2024.10062469

Article History:

Received:	21 November 2023
Last revised:	07 December 2023
Accepted:	12 December 2023
Published online:	11 April 2024

Research on scoring mechanism of spoken English self-study system taking into account artificial intelligence technology and speech knowledge recognition algorithm

Ning Li

Public Teaching Department,
Henan Vocational College of Tuina,
Luoyang, Henan, 471023, China
Email: ln19802023@163.com

Abstract: With the development of computers and the continuous progress of speech recognition technology, the use of computer-assisted language learning (CALL) system for self-study of English has become an important way for students to learn English. This paper proposes an advanced scoring mechanism for spoken English self-learning system incorporating artificial intelligence technology and speech knowledge recognition algorithm. This new mechanism aims to use intelligent technology to more accurately assess students' oral expression skills and provide them with more personalised and targeted learning support.

Keywords: speech technology; Language learning system; scoring mechanism.

Reference to this paper should be made as follows: Li, N. (2024) 'Research on scoring mechanism of spoken English self-study system taking into account artificial intelligence technology and speech knowledge recognition algorithm', *Int. J. Information and Communication Technology*, Vol. 24, No. 3, pp.350–365.

Biographical notes: Ning Li studied in Henan University from 2001–2003 and received her Bachelor's degree in 2003. She has been working as an English teacher in Henan Vocational College of Tuina since graduation in 2003. She has published a total of ten papers in Chinese journals. Her major area of study includes English teaching and research, English Translation and English language research.

1 Introduction

The development of computer technology provides an opportunity for the innovation of English teaching methods. Computer-assisted language (CALL) has become an inevitable trend in the reform of English language learning, and it will also become a new teaching method in the information age of colleges and universities. The superb application of human-computer interaction technology and voice processing technology in CALL gives learners more learning opportunities to enter a more diversified simulation learning environment. CALL has become the best way to learn oral English. The core technology of the CALL system is the voice scoring mechanism. The voice quality assessment is an accurate assessment of the learners' pronunciation practice through a computer that applies voice processing technology, not an expert assessment. The current research on the scoring mechanism of the CALL system is mostly focused on extracting the acoustic characteristics of the voice signal, which ignores the voice signal hidden in other aspects. Therefore, the use of better scoring methods and scoring mechanisms can more accurately evaluate the learner's voice, which not only helps learners find their own pronunciation insufficient, but also better promotes the efficiency of oral English learning.

2 Research trends at home and abroad based on the automatic evaluation method of spoken English based on deep learning technology

At present, artificial intelligence has made great progress in speech recognition, which makes the speech recognition ability of artificial intelligence comparable to that of humans. Before the advent of deep learning, the evaluation process usually involves language test participants arbitrarily selecting speech samples from the speech library, and the evaluation process needs to follow the correct pronunciation of these samples. The system extracts the feature information of the two separately during operation for comparison. It evaluates the pronunciation quality by calculating the Euclidean distance between the standard and the sample performance parameters, but deep learning changes every link of this process.

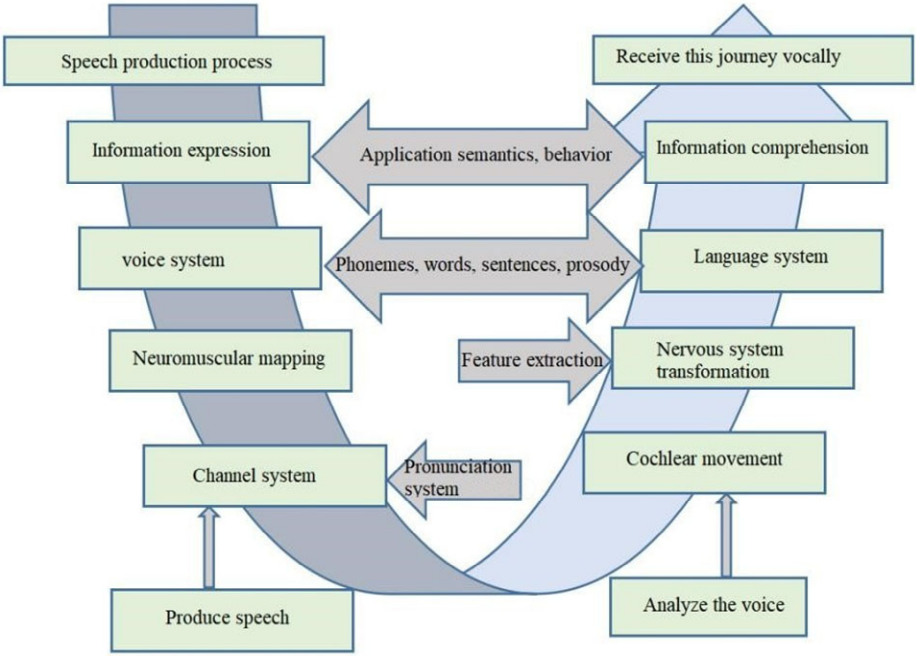
Voice evaluation is a complex nonlinear dynamic process. The development of speech assessment does not follow a certain static formula, it can get the result by inputting structural parameters. In the dynamic interaction, it gradually evolves into the output we observe. Although the deep neural network model is the dominant structure of the speech recognition audio model, the result of speech recognition determines everything afterwards, so the automatic evaluation of pronunciation quality should have the possibility of its own independent path (Zechner et al., 2009; Byun and Lee, 2021). The scoring mechanism is based on the principles of artificial intelligence, natural language processing, and speech recognition.

3 Speech knowledge recognition

3.1 Mechanism of voice generation and reception

This section details the implementation of speech recognition algorithm in evaluating spoken English proficiency. Humans can exchange information with each other through words, so how do words come into being? Speech is composed of a series of phonemes, which are the sounds that make up the language. The first stage of artificially generated speech is to decide what content they want to send to each other, and then translate it into language (Sutrisno, 2018). The brain sends motor neuron instructions to the vocal cords, and various muscles of the vocal cords move and vibrate in the air to form voice waves. After the voice signal is generated and the voice information is transmitted to the listener, the voice perception process begins. This process is shown in Figure 1.

Figure 1 Voice signal generation and reception (see online version for colours)


Speech production process
Receive this journey vocally
Information expression
Application semantics, behavior
Information comprehension
voice system
Phonemes, words, sentences, prosody
Language system
Neuromuscular mapping
Feature extraction
Nervous system transformation
Channel system
Pronunciation system
Cochlear movement
Produce speech
Analyze the voice

3.2 *Acoustic model and Language model of speech*

1 *Acoustic model.* Acoustic modelling is an important part of the speech recognition system. Its goal is to provide an effective method to calculate the distance between the speech feature vector sequence and each pronunciation model. The design of the acoustic model is closely related to the characteristics of language pronunciation.

First of all, the sound model needs to select basic acoustic units. Acoustic units can be divided into phoneme units, syllable (word) units, semi-syllable units, and so on. The choice of basic acoustic unit has a greater impact on the amount of speech training data and the speed of speech recognition. The larger the selection of the basic acoustic unit, the easier it is to include the matching phenomenon in the model, which is useful for improving the recognition speed of the system, but this increases the amount of computer calculations, model storage requirements and training data. Choosing a smaller acoustic unit requires relatively little training data, but it is difficult to locate and segment the speech accordingly (Laptev et al., 2021).

Therefore, in the process of designing the acoustic model, the relevant personnel must fully consider the actual situation and requirements of the system and select the appropriate acoustic unit. At the same time, they must improve the existing deficiencies by other means.

2 *Language model.* As we all know, a sequence formed by randomly selecting some words from the vocabulary may not form a natural language sentence. Only a sentence that conforms to the grammar is considered a sentence. Language models are divided into grammar-based and statistical-based language models (Hu and Zhao, 2021).

A grammar-based language model is to summarise grammatical rules and even semantic rules. When statistical language modelling, a large number of text documents are used to count the probability of each word and its associated conditional probability, and this knowledge is combined with the model's matching voice. The results are evaluated. Suppose $W = w_1, w_2, \cdots w_Q$, the probability can be expressed as shown in formula (1):

$$P(W) = P\left(w_1, w_2, \cdots w_Q\right) = P\left(w_1\right) P\left(w_2 \middle| w_1\right) P\left(w_3 \middle| w_1 w_2\right) \cdots P\left(w_Q \middle| w_1 w_2 \cdots w_{Q-1}\right) \quad (1)$$

3.3 *Principles of speech knowledge recognition*

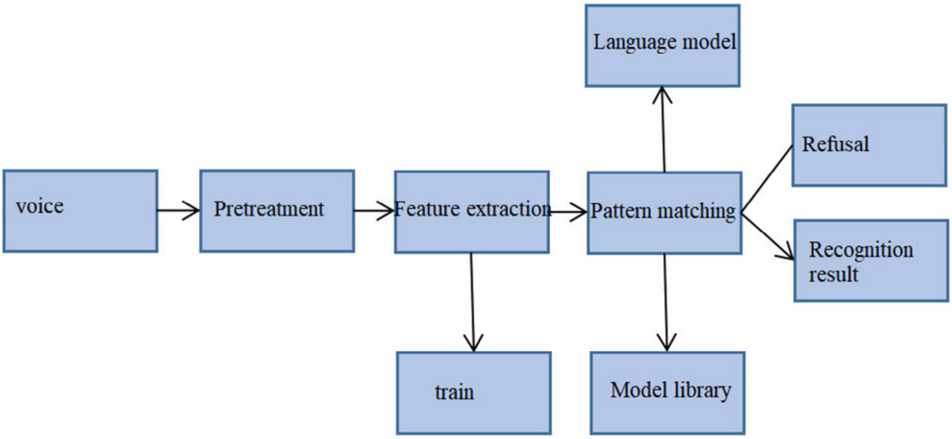
Speech recognition is essentially a pattern recognition process, the basic process is (Rajesh Kumar et al., 2021):

The recognition system establishes a search grid according to the given grammar. Collect standard voice signals, then perform noise reduction processing on the signals, extract voice features for model training, and build an audio model library.

Allocate the extracted speech features to the decoder for processing. The decoder input according to the voice feature results, and finds the most matching result in the search space.

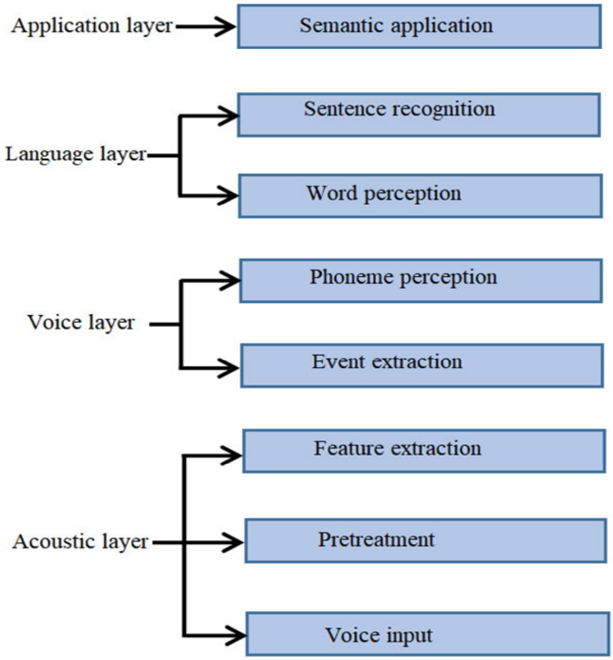
Figure 2 is the principle block diagram of the automatic speech recognition system.

Figure 2 The principle block diagram of the speech recognition system (see online version for colours)



According to the principle flow of speech recognition in Figure 1 and Figure 2 of speech generation and understanding in the previous section. A hierarchical model of speech recognition can be obtained, as shown in Figure 3.

Figure 3 Speech recognition flowchart (see online version for colours)



3.4 Problems and improvements faced by speech recognition

Due to the differences in languages in different regions, it is difficult to perform standard input for speech recognition, which leads to a high rate of misrecognition in actual use. Even if the language is the same, everyone's pronunciation habits are not the same, which also results in a lack of standardisation and uniformity in the speech recognition system. Therefore, first, users can perform adaptive training on the speech recognition system, and they can familiarise the system with their own pronunciation and improve the recognition accuracy.

Environmental noise has a great interference effect on the accuracy of speech recognition. Therefore, when the speech recognition function is developed, devices with better noise reduction performance for speech input are added to reduce the noise in the speech input. Secondly, when extracting speech signal features mixed with noise, select feature parameters with strong anti-interference ability; finally, when the speech recognition system is training, it is necessary to fully consider the problem of noise interference, and conduct targeted training to improve the robustness of system recognition.

When people communicate, they rarely pronounce one word in isolation. In most cases, they pronounce it continuously according to their own habits. This makes the originally isolated acoustic unit affected by the context and blurs and mutates. Therefore, no matter what modelling units (words, syllables, consonants, vowels, phonemes) are selected in the speech recognition system, the interaction between these units needs to be refined.

3.5 Voice scoring mechanism and scoring technology

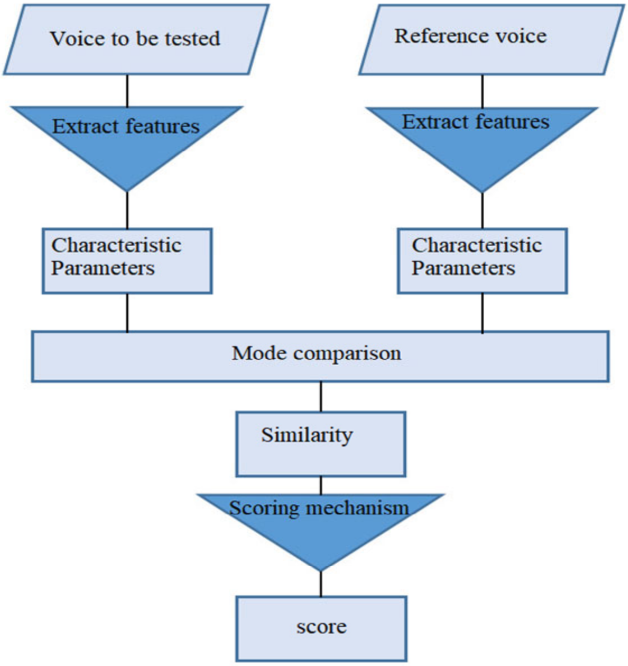
The scoring mechanism is the core technology of the self-learning system of spoken English. The main purpose is to use a computer to automatically assess whether a person's English pronunciation is correct, list the similarities and differences in the form of charts, use audio or animation to prompt the correct pronunciation, and provide pronunciation or text information pronunciation suggestions, so that learners can improve through repeated practice the pronunciation of personal spoken English.

At present, the oral English learning system can be divided into two categories according to the voice scoring technology used. One is a scoring method based on the comparison of voice features, which evaluates a piece of voice from a more subjective perspective, this is generally implemented by dynamic time warping (DTW) technology. The other is the scoring method based on acoustic model. This method is more objective, mainly based on Hidden Markov Model (HMM) technology.

3.6 Voice score based on feature comparison

Voice grading is based on feature comparison. It mainly compares the reference standard voice and the learner's pronunciation (Bedward et al., 1991). The specific process is shown in Figure 4.

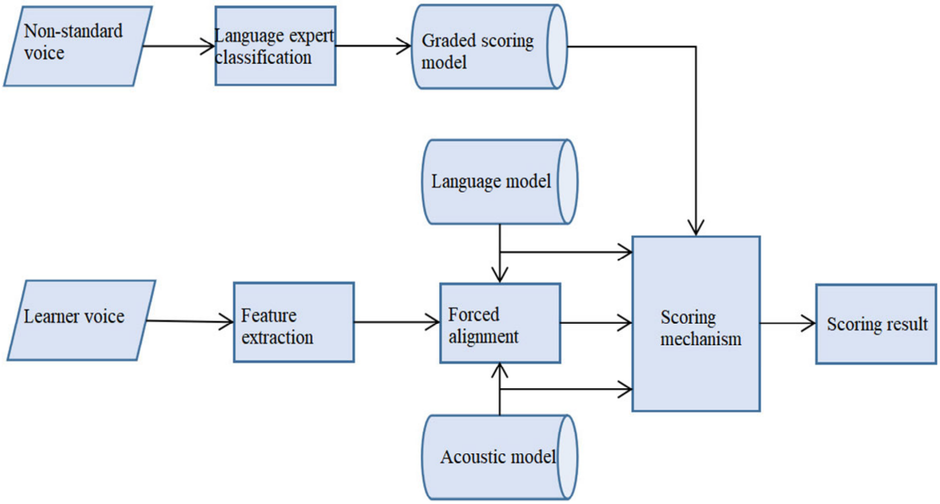
Figure 4 DTW-based voice scoring process (see online version for colours)



3.7 Voice scoring based on acoustic model

The voice scoring based on the acoustic model does not need to refer to the standard voice, and the process is shown in Figure 5. First, non-standard voices need to be collected and scored by language experts, thus forming a grading scoring model. Then extract the learner's voice features, use the predefined language model and the trained acoustic model to align, and obtain the scoring result through an appropriate scoring mechanism (Laukka et al., 2016).

Figure 5 HMM-based voice scoring process (see online version for colours)



4 Speech feature parameter extraction

4.1 Linear prediction cepstrum coefficient LPCC

In the speech recognition system, in order to improve the stability of the parameters, the linear prediction coefficient LPC is rarely used directly, but the method of isomorphic signal processing is used to express the cepstrum-linear prediction coefficient LPCC in the domain. Statistical analysis methods are utilised to validate the accuracy of parameter extraction.

4.1.1 Homomorphic processing

The modulation signal of audio and video signals is a nonlinear product signal or a convolution signal. The concept of isotropic signal processing is to convert these nonlinear signals into linear signals for processing. Homomorphic processing can be divided into product homomorphic processing and convolutional homomorphic processing according to the classification of its processing signals (Ozyapici and Bilgehan, 2016). An important theory of homomorphic processing is that any homomorphic system can be expressed as a cascade of three homomorphic systems and satisfies the superposition theorem. Three steps are required to convert the convolution signal into an additional signal in the time domain.

The first step is to convert the convolution signal into a product signal through z-transformation, as shown in formula (2):

$$Z[x(n)] = X(z) = X_1(z) \bullet X_2(z) \tag{2}$$

The second step is to use logarithmic operation to convert the product signal into an addition signal, as shown in formula (3):

$$\ln X(z) = \ln[X_1(z) \bullet X_2(z)] = \ln X_1(z) + \ln X_2(z) = \hat{X}_1(z) + \hat{X}_2(z) = \hat{X}(z) \tag{3}$$

The third step is to perform the inverse *Z* transform on the added signal obtained in the previous step, and restore it to the time domain, as shown in formula (4):

$$Z^{-1}\left[\hat{X}_1(z) \bullet \hat{X}_2(z)\right] = \hat{X}_1(n) + \hat{X}_2(n) = \hat{X}(n) \tag{4}$$

After the obtained time-domain signal $x(n)$ is linearly processed by the linear system in the homomorphic system, the characteristic system can be used to restore it to a convolutional signal.

4.1.2 Cepstrum and cepstrum solution

Theoretically, the time-domain sequence $\hat{x}(n)$ is called the ‘complex cepstrum’ of the time-domain sequence $x(n)$. In most signal processing, the convergence range of $\hat{x}(z)$ and $X(z)$ includes the unit circle. Therefore, the discrete Fourier transform (DFT) can be used to replace the *Z* transform in equation (2) (Perracchione et al., 2021). The inverse transform replaces the *Z* transform in formula (4), so that formula (2) to (4) can be rewritten as shown in formula (5) to (7):

$$F[x(n)] = X\left(e^{jw}\right) \tag{5}$$

$$\hat{X}\left(e^{jw}\right) = \ln\left[X\left(e^{jw}\right)\right] \tag{6}$$

$$\hat{X}(n) = F^{-1}\left[\hat{X}\left(e^{jw}\right)\right] \tag{7}$$

The definition of ‘cepstrum’ is similar to the definition of ‘complex cepstrum’, the difference is that the logarithmic operation in the above formula is replaced by the logarithmic operation of the modulus. If $c(n)$ represents cepstrum, the definition of cepstrum is formula (8):

$$c(n) = F^{-1}\left[\ln\left|X\left(e^{jw}\right)\right]\right. \tag{8}$$

It can be seen from equation (8) that the ‘cepstrum’ is actually the inverse Fourier transform of the log-magnitude spectrum of the $x(n)$ series. The cepstrum coefficient is a good speech feature parameter, it can eliminate the stimulus information in the human voice process, and highlight the characteristics of the vocal tract response. It has been widely used in speech recognition and achieved good results. The ‘cepstrum’ solution process is as follows:

When the sequence $x(n)$ meets the minimum phase, $X(z)$ is analytic in the unit circle. At this time, the recurrence relation can be used to simplify the calculation, as shown in formula (9):

$$\hat{X}(z) = \frac{d \log X(z)}{dz} = \frac{X'(z)}{X(z)} \tag{9}$$

Multiply both sides of the above formula by $zX(z)$, and then perform the inverse Z transformation to obtain formula (10):

$$nx(n) = \sum_{-\infty}^{\infty} k\hat{x}(k)x(n-k) \tag{10}$$

When n is not equal to 0, then both sides are divided by n, and formula (11) is obtained:

$$x(n) = \sum_{-\infty}^{\infty} \frac{k}{n}\hat{x}(k)x(n-k), n \neq 0 \tag{11}$$

When $n < 0$, the condition is satisfied: $x(n) = 0$, so formula (12) is obtained:

$$x(n) = \begin{cases} 0, n < 0 \\ \sum_{-\infty}^{\infty} \frac{k}{n}\hat{x}(k)x(n-k), n > 0 \end{cases} \tag{12}$$

Therefore, the 'cepstrum' recurrence formula is obtained as formula (13):

$$x(n) = \begin{cases} 0, n < 0 \\ \frac{x(n)}{x(0)} \sum_{k=0}^{n=1} \frac{k}{n}\hat{x}(k)\frac{x(n-k)}{x(0)}, n > 0 \end{cases} \tag{13}$$

4.1.3 LPCC solution

LPCC or linear prediction cepstrum coefficient is a very important speech feature parameter. Unlike the above-mentioned 'complex cepstrum' and 'cepstrum', the cepstrum of LPCC and LPC is directly obtained (Lee et al., 2015). The recurrence relationship between LPCC and LPC is shown in formula (14) to (16):

$$c_0 = \log G^2 \tag{14}$$

$$c_n = a_n + \sum_{k=1}^{n-1} \frac{k}{n} c_k a_{n-k}, (1 \leq n \leq N) \tag{15}$$

$$c_n = \sum_{k=1}^{n-1} \frac{k}{n} c_k a_{n-k}, (n > N) \tag{16}$$

4.2 Mel cepstrum coefficient MFCC parameter extraction

In practical applications, the following steps are usually used to calculate MFCC parameters:

First, perform pre-emphasis, windowing, and framing on the original voice signal, and determine the number of sampling points of each frame of the signal during framing, that is, the length *N* of each frame of the voice signal. Suppose that the framed frame sequence is transformed from the time domain *s*(*n*) to the frequency domain through FFT, and then the frequency spectrum is obtained, and the short-term energy spectrum *S*(*n*) is obtained by calculating the modulus square (Abeysinghe et al., 2021).

Convert the triangular bandpass filter in the linear time domain to the Mel frequency domain to obtain the Mel frequency filter bank *Hm*(*k*). The design of the Mel frequency filter bank *Hm*(*k*) is as shown in formula (17):

$$H_m(k) = \begin{cases} 0, k < f_{m-1} ork > f_{m+1} \\ \dfrac{k - f_{m-1}}{f_m - f_{m-1}}, f_{m-1} < k < f_m, m = 0,1,2,\cdots,M-1 \\ \dfrac{f_{m+1} - k}{f_{m+1} - f_m}, f_m < k < f_{m+1} \end{cases} \tag{17}$$

Calculate the sum of the products of *S*(*n*) and *Hm*(*n*) at *M* discrete frequency points to obtain the energy spectrum *S*(*n*). The output Pm after passing through the Mel band-pass filter bank *Hm*(*k*), $0 \le m \le M - 1$, the number of parameters *Pm* is *M*, and the calculation is shown in formula (18) (Hussain et al., 2020):

$$P_m = \sum_{k=f_{m-1}}^{f_{m+1}} H_m(k) S_n(k), m = 0,1,2,\cdots,M-1 \tag{18}$$

Perform logarithmic operation on *Pm* to get the parameter *Lm*, where *m* = 0,1,2,...,*M* – 1, and then transform the result to the cepstrum domain by doing discrete cosine transform (DCT) on *Lm*, and get MFCC coefficient *Cmel*, the process is shown in formula (19):

$$\begin{aligned} &L_m = \ln(P_m), m = 0,1,2,\cdots,M-1 \\ &C_{mel}(k) = \sum_{m=0}^{M-1} L_m \cos[k(m-0.5)\pi / M], k = 1,2,3,\cdots,N \end{aligned} \tag{19}$$

The standard Mel cepstrum parameters only reflect the static characteristics of the speech signal, and the speech signal actually contains many dynamic characteristics that are more sensitive to the human ear. Therefore, in actual recognition, it is necessary to use differential cepstrum parameters to describe this dynamic characteristic of the speech signal, and its calculation is shown in formula (20):

$$d(n) = \frac{1}{\sqrt{\sum_{i=-k}^{k} i^2}} \sum_{i=-k}^{k} i \bullet c(n+i), 1 \le n \le N \tag{20}$$

Compared with LPCC parameters, MFCC parameters have a significant improvement in recognition performance, but there are still certain defects. That is, the dimensional components of the MFCC parameters have different characterisation functions for speech features, and their contributions to the final recognition result are also different. Among them, components with a small contribution rate cannot improve the recognition rate. Therefore, in this article, we propose an improvement plan. The specific problems and improvement plans are as follows:

A number of experimental results show that the characterisation ability of each component in the MFCC parameters and its contribution to the recognition result are different, and the mean variance between the components of each dimension is very different. Therefore, each dimension component of an MFCC parameter can be multiplied by its weighting factor, and the weighted MFCC parameter is represented by Weighing-MFCC (WMFCC).

MFCC parameters can only reflect the static characteristics of the voice, but these cannot reflect the dynamic characteristics. Therefore, firstly, ΔWMFCC is used to represent the primary difference that can reflect the rate of change of the characteristic parameter components, and secondly, A2WMFCC is used to represent the secondary difference that reflects the acceleration of the parameter component change. Finally, the WMFCC parameters and its first-order and second-order differences are combined to form a new feature parameter, which allows the feature parameter to reflect the static characteristics of the speech and its dynamic characteristics.

After the dimensionality of the feature parameters is changed from 14 to 42 dimensions, the amount of calculation and storage of data increases rapidly, which brings a lot of calculation and storage pressure to the recognition system. Therefore, the data storage and calculation complexity in the subsequent processing of speech recognition can be reduced by dimensionality reduction, and the parameters can be optimised by the principal component analysis method to improve the efficiency of training and recognition.

5 Speech knowledge recognition algorithm

5.1 DTW algorithm

The first recognition method used in speech recognition is the template matching method. Template matching method is one of the most commonly used methods to calculate similarity. In template matching, the feature parameters extracted through training are first formed into a template, and then in the recognition step, the similarity between each template and the feature parameters of the pattern to be recognised is calculated, and finally the template type is determined. However, there are still various problems in the discrimination process. Even if the same speaker pronounces the same word, there may be differences in the obtained voice samples. The most common is the inconsistent pronunciation length. In order to solve this problem, DTW algorithm is proposed (Jin et al., 2020). The DTW algorithm is implemented based on the nearest neighbour principle, which is currently the best nonlinear time warping template matching algorithm. The specific steps of the DTW algorithm are:

Initialise: Let $m_i = i, i = 1, 2, \cdots, M, n_1 = 1, n_M = N$ define the path function as $n_i = \Phi(m_i)$, and initialise the function $\Phi(0) = 0$, $\Phi(1) = 1$.

Find the values of $D[(m_{ii}, n_i)]$ and $\Phi(m_i)$ when $i = 2,3,...M$.

Get the total distortion of the path by $D[(M,N)]$.

Recursively from the point (M, N) where $i = M$ according to the above method, the best path can be obtained, where i goes from M to 0.

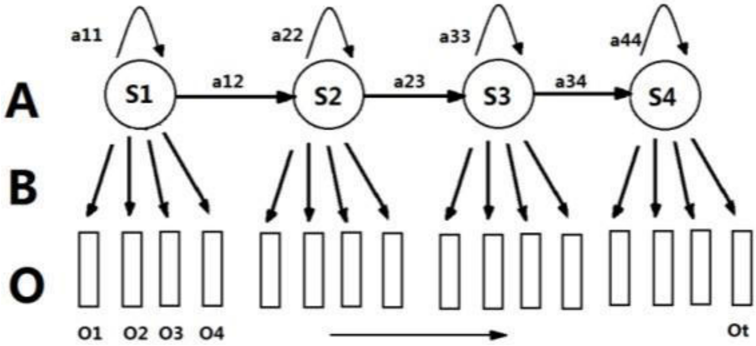
5.2 Hidden Markov model

The Markov model is a statistical model proposed by Andrei Markov (1856–1922) with a wide range of applications. Researchers have developed many variants on the basis of the Markov model, and HMM is one of them. The successful application of HMM in speech recognition has completely changed the history of speech recognition and this has had a profound impact.

The state of the Markov chain is sent randomly. If the outside world cannot observe the state Xt at any time t, and can only get an arbitrary output observation vector, and the observation vector is only related to the current state, then this state hidden Markov chain is called a hidden Markov model, or simply HMM model. Among them, the observation vector can be continuously distributed or discretely distributed (Roy et al., 2021).

According to the definition of HMM, the HMM model is a double random process: on the one hand, the state transition of the model is random, and on the other hand, the output of the state is also random. All these random processes are reflected in the model parameters A and B. As shown in Figure 6.

Figure 6 HMM model state representation diagram



In order to apply HMM to speech recognition, three problems need to be solved. The following are the problems to be solved and their solutions:

Question 1: The calculation of output probability.

Solution-forward and backward algorithm.

This article explains the forward algorithm as an example: first define the forward probability vector: $\alpha_t(i) = P(o_1, o_2, \cdots, o_t, q_t = \theta_i \mid \lambda)$, which means that under the premise of a given HMM model λ, the part of the observation sequence output from time 1 to time *t* is $\{o_1, o_2, \cdots, o_t\}$, and it is at time *t* Probability of state 0. Then use the defined forward vector to calculate the output conditional probability, the steps are as follows:

Then use the defined forward vector $\alpha_t(i)$ to calculate the output conditional probability $P(O \mid \lambda)$, the steps are as follows:

Step 1 Initialisation. When $1 \le i < N$, formula (21) is obtained:

$$\alpha_t(i) = \pi_i b_i(o_1) \tag{21}$$

Step 2 Recursive calculation. For all $1 \le t < T - 1; 1 \le j \le N$, the formula (22) is obtained:

$$\alpha_{t+1}(j) = \left[\sum_{i=1}^{N} \alpha_t(i) a_{ij}\right] b_j(o_{t+1}) \tag{22}$$

Step 3 Terminate the calculation, the calculation formula is (23):

$$P(O \mid \lambda) = \sum_{i=1}^{N} \alpha_T(i) \tag{23}$$

Question 2: Determination of the best state sequence

Solution-Viterbi algorithm.

First define as shown in formula (24):

$$\delta_t(i) = \max_{q_1 q_2 \cdots q_{t-1}} P\left(q_1 q_2 \cdots q_{t-1}, q_t = \theta_i, o_1 o_2 \cdots o_t \middle| \lambda\right) \tag{24}$$

Then the steps to use $\delta_t(i)$ to find the best state sequence are as follows:

Step 1 Initialise when $1 \le i \le N$, get formula (25):

$$\delta_t(i) = \pi_i b_i(o_1) \tag{25}$$

Step 2 Recursive calculation. For all $2 < t \le T$; $1 < j \le N$, the formula (26):

$$\begin{cases} \delta_t(j) = \max\limits_{1 \le i \le N} \left[\delta_{t-1}(i) a_{ij}\right] b_j(o_t) \\ \varphi_t(j) = \arg\max\limits_{1 \le i \le N} \left[\delta_{t-1}(i) a_{ij}\right] \end{cases} \tag{26}$$

Step 3 Terminate the calculation, the calculation formula is (27):

$$\begin{cases} P^* = \max\limits_{1 \le i \le N} \left[\delta_T(i)\right] \\ q_t^* = \arg\max\limits_{1 \le i \le N} \left[\delta_T(i)\right] \end{cases} \tag{27}$$

Step 4 Find the state sequence, the best state sequence is shown in formula (28):

$$q_t^* = \varphi_{t+1}\left(q_{t+1}^*\right), t = T - 1, T - 2, \cdots, 1 \tag{28}$$

Question 3: Adjustment and optimisation of HMM model parameters

Solution-Baum-Welch algorithm.

First define $\varepsilon_t(i, j)$. Given the model λ and the training sequence O, the Markov chain of the HMM model is in state i at time t, and the probability of being in state j at time t+1. Its expression is shown in formula (29):

$$\varepsilon_t(i,j) = P\left(q_t = i, q_{t+1} = j \mid O, \lambda\right) \tag{29}$$

Derive the formula (30):

$$\varepsilon_t(i,j) = \frac{\alpha_t(i) a_{ij} b_j \beta_{t+1}(i)}{P(O \mid \lambda)} \tag{30}$$

Then, the probability that the Markov chain of the model is in state i at time t is formula (31):

$$\varepsilon_t(i) = P\left(q_t = \theta, O \mid \lambda\right) = \sum_{j=1}^{N} \varepsilon_t(i,j) = \frac{\alpha_t(i)\beta_t(i)}{P(O \mid \lambda)} \tag{31}$$

Although HMM model technology is a good speech recognition technology in theory, there are still some practical problems in the specific application process. The following are related problems and solutions:

The research shows that the selection of the number of states in the HMM model is not the better. Therefore, relevant researchers have developed specific control strategies: For English, the number of states selected in the HMM model is 6 when the effect is best; For Chinese word recognition, the number of HMM states is between 6–9.

In the application of speech recognition, the two-transition structure of the HMM model is often selected, that is, the state can only transition to itself or the next state. Therefore, when calculating the path of the Viterbi algorithm to search for the best state sequence, there are only two alternative paths for each calculation, which results in the Viterbi algorithm having no advantage compared with the full probability algorithm. It is possible to take the logarithm of the probability result in the algorithm, and convert a large number of multiplication operations into addition operations, which will greatly reduce the amount of operation of the entire algorithm.

In addition to the small probability continuous multiplication in the Viterbi algorithm, which can easily lead to data overflow, a large number of continuous multiplications are also performed in the calculation process of the Baum-Welch algorithm. Therefore, this will also cause data to overflow. In response to this problem, related researchers have proposed a normalised solution, which can effectively solve the problem of data overflow in the actual application process.

When the Baum-Welch algorithm selects the initial model, different initial values may cause different training results. If the initial model is not properly selected or the representation is not accurate enough, it may lead to too many iterations or failure to converge. So it is very important to initialise the HMM model scientifically and reasonably. At present, the random or average method can be directly used to select the initial value that can meet the probability requirements. As for the setting of output probability B, enough attention should be paid to it. Generally, it is estimated based on existing experience. If necessary, it can be set multiple times. At the same time, Viterbi algorithm is used for multiple verifications until the requirements are met. Background noise and variability in accents pose significant challenges to developing an accurate automated speech scoring system. The automated scoring results demonstrate over 90% alignment with human expert evaluations across fluency, pronunciation, grammar and vocabulary assessment dimensions.

6 Conclusions

At present, the research field of the CALL system based on voice technology is still in its infancy, and voice recognition is also a hot research direction of voice technology. This article is exploratory to improve the performance of the scoring mechanism in the CALL system. There are still many problems to be solved. In the future, more efforts are needed to improve the scoring mechanism of the CALL system.

In addition, the perfect function of the scoring mechanism in the CALL system is to enable oral learners to accurately understand their own learning conditions, which can help learners detect and locate pronunciation errors, and then give corrective suggestions for learners' learning conditions. Therefore, the judgment and correction of pronunciation errors are also an important part of the scoring mechanism. Accurately discovering and locating errors and giving suggestions for improvement are the direction of future efforts. Potential solutions involve integrating speech enhancement techniques to improve signal clarity. Further research should focus on robust algorithms that can adapt to diverse accents and languages.

References

Abeysinghe, A., Fard, M., Jazar, R., Zambetta, F. and Davy, J. (2021) 'Mel frequency cepstral coefficient temporal feature integration for classifying squeak and rattle noise', *The Journal of the Acoustical Society of America*, Vol. 150, No. 1, pp.228–342.

Bedward, M., Pressey, R.L. and Nicholls, A.O. (1991) 'Scores and score classes for evaluation criteria: a comparison based on the cost of reserving all natural features', *Elsevier*, Vol. 56, No. 3, pp.231–290.

Byun, S.W. and Lee, S.P. (2021) 'A study on a speech emotion recognition system with effective acoustic features using deep learning algorithms', *Applied Sciences*, Vol. 11, No. 4, pp.234–452.

Hu, G. and Zhao, Q. (2021) 'Multi-model fusion framework based on multi-input cross-language emotional speech recognition', *International Journal of Wireless and Mobile Computing*, Vol. 20, No. 1, pp.563–782.

Hussain, T., Iqbal, N., Maqbool, H.F., Khan, M. and Tahir, M. (2020) 'Amputee walking mode recognition based on mel frequency cepstral coefficients using surface electromyography sensor', *International Journal of Sensor Networks*, Vol. 32, No. 3, pp.142–273.

Jin, D., Li, R. and Xu, J. (2020) 'Multiscale community detection in functional brain networks constructed using dynamic time warping', *IEEE Transactions on Neural Systems and Rehabilitation Engineering: A Publication of the IEEE Engineering in Medicine and Biology Society*, Vol. 28, No. 1, pp.190–340.

Laptev, A., Andrusenko, A., Podluzhny, I., Mitrofanov, A., Medennikov, I. and Matveev, Y. (2021) 'Dynamic acoustic unit augmentation with bpe-dropout for low-resource end-to-end speech recognition', *Sensors*, Vol. 21, No. 9, pp.45–79.

Laukka, P., Elfenbein, H.A., Thingujam, N.S., Rockstuhl, T., Iraki, F.K., Chui, W. and Althoff, J. (2016) 'The expression and recognition of emotions in the voice across five nations: a lens model analysis based on acoustic features', *Journal of Personality and Social Psychology*, Vol. 111, No. 5, pp.54–92.

Lee, Y-C., Pang, J-S. and Mitchell, J.E. (2015) 'An algorithm for global solution to bi-parametric linear complementarity constrained linear programs', *Journal of Global Optimization*, Vol. 62, No. 2, pp.112–443.

Ozyapici, A. and Bilgehan, B. (2016) 'Finite product representation via multiplicative calculus and its applications to exponential signal processing', *Numerical Algorithms*, Vol. 71, No. 2, pp.23–45.

Perracchione, E., Massone, A.M. and Piana, M. (2021) 'Feature augmentation for the inversion of the Fourier transform with limited data', *Inverse Problems*, Vol. 37, No. 10, pp.43–82.

Rajesh Kumar, T., Vijendra Babu, D., Malarvezhi, P., Velu, C.M., Haritha, D. and Karthikeyan, C. (2021) 'Boltzmann–Dirichlet process mixture: a mathematical model for speech recognition', *Journal of Physics: Conference Series*, Vol. 1964, No. 4, pp.103–200.

Roy, P.P., Kumar, P. and Kim, B.G. (2021) 'An efficient sign language recognition (SLR) system using camshift tracker and Hidden Markov Model (HMM)', *SN Computer Science*, Vol. 2, No. 2, pp.230–492.

Sutrisno, A. (2018) 'Problems of speech perception experienced by the EFL learners', *Theory and Practice in Language Studies*, Vol. 8, No. 1, pp.124–231.

Zechner, K., Higgins, D., Xi, X. and Williamson, D.M. (2009) 'Automatic scoring of non-native spontaneous speech in tests of spoken English', *Speech Communication*, Vol. 51, No. 10, pp.1422–1476.