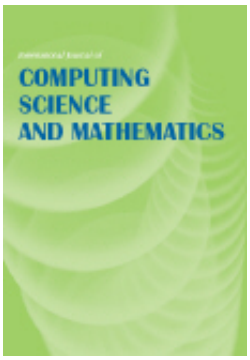


International Journal of Computing Science and Mathematics

ISSN online: 1752-5063 - ISSN print: 1752-5055
https://www.inderscience.com/ijcsm

A method for capturing English oral pronunciation errors based on speech recognition

Wenna Dou

DOI: 10.1504/IJCSM.2024.10068571

Article History:

Received:	12 March 2024
Last revised:	30 July 2024
Accepted:	19 October 2024
Published online:	06 May 2025

A method for capturing English oral pronunciation errors based on speech recognition

Wenna Dou

College of Humanities,
University of Civil Engineering and Architecture,
Beijing 102600, China
Email: douwennavip@163.com

Abstract: To improve the accuracy and timeliness of capturing pronunciation errors, the paper proposes a new method for capturing English oral pronunciation errors based on the speech recognition process. Using a voice production system to collect raw English spoken pronunciation signals and extract the features of the speech signals. Then, after determining the confidence level of the intonation points, hidden Markov model (HMM) classification algorithm is used to classify the intonation points and establish a spoken pronunciation comparison database containing standard state sequences. Finally, the degree component signal detection method is used to determine the spectral features of pronunciation errors. By comparing the spectral features with standard state sequences, incorrect English spoken pronunciation is captured. Experiment shows that the recognition accuracy of this method remains above 97%, and the maximum accuracy of capturing pronunciation errors can reach 98.74%. The capture time remains within 3s, indicating that this method has achieved the design expectations.

Keywords: speech recognition; English speaking test; capture pronunciation errors; classification of intonation points; standard state sequence.

Reference to this paper should be made as follows: Dou, W. (2025) 'A method for capturing English oral pronunciation errors based on speech recognition', *Int. J. Computing Science and Mathematics*, Vol. 21, No. 1, pp.32–47.

Biographical notes: Wenna Dou received her Master's Degree in Foreign Linguistics and Applied Linguistics from Capital Normal University in 2012. Currently she is an Assistant Professor in the College of Humanities of Beijing University of Civil Engineering and Architecture. Her research interests include CALL, discourse analysis and cross-cultural communication.

1 Introduction

With the rapid development of economic globalisation, people's frequency of using English in work and social life is gradually increasing. The level of spoken English pronunciation has a direct impact on the quality of English communication (Lei et al., 2021; Yu et al., 2022). Due to the fundamental differences in pronunciation rules between English and Chinese, there are certain problems with the pronunciation of spoken English

by Chinese people. The subtle differences in pronunciation cannot be distinguished well, resulting in a large number of difficult to understand oral phenomena and inaccurate pronunciation in English. Therefore, it is necessary to conduct high-quality detection and capture of English spoken pronunciation, in order to improve the standardisation level of English spoken language (Wang et al., 2023).

With the continuous optimisation of speech signal processing technology, people can achieve English spoken pronunciation error recognition by constructing an expert evaluation database. Based on this principle, some experts and scholars have proposed corresponding methods for capturing English oral pronunciation errors. Wang et al. (2021) proposed a detection method that integrates speech action features and acoustic features. This method analyses the difference between abnormal voice and normal voice, extracts two kinds of voice action features: displacement and speed, extracts three acoustic features: Mel cepstrum coefficient, fundamental frequency and Formant, normalises the two types of features, uses Kernel principal component analysis to reduce dimensions, and then uses support vector machine to recognise the wrong voice features. However, if the normalisation process is inaccurate or unreasonable, it can lead to problems in comparing and classifying different features, thereby affecting recognition accuracy. Huo and Xie (2022) takes pronunciation attribute as the core content, carries out modelling analysis on fine Granularity pronunciation attribute, proposes a method of modelling fine Granularity pronunciation attribute (FSA), and tests it in cross language attribute recognition and pronunciation error detection. However, this method focuses on fine-grained pronunciation attributes for modelling and analysis. If the selected fine-grained attribute model is insufficient to cover all possible pronunciation error situations, it will result in low accuracy in pronunciation error capture. Lv et al. (2021) proposed a method for detecting spoken language pronunciation by integrating language models. After building a language model from pinyin to Chinese characters, the speech frame decomposition model was designed to connect the output of the Acoustic model and the input of the language model, which overcame the problem of poor integration of language models and Acoustic model, and further reduced the error rate of detection results. Although this method combines language and acoustic models, acoustic models may have certain limitations in capturing and identifying pronunciation errors. Acoustic models may require more speech samples and more comprehensive acoustic features to capture and distinguish different pronunciation errors, increasing the duration of pronunciation error capture.

Based on the above analysis, it can be found that traditional methods do not perform well in terms of accuracy in oral speech recognition, accuracy in capturing pronunciation errors, and duration of pronunciation error capture. In response to these situations, this study applies speech recognition technology to propose a new method for capturing English oral pronunciation errors. The specific design concept of this method is as follows:

1 Using a speech production system to collect raw English spoken pronunciation signals, based on matching standardised speech signals, the Mel frequency cepstrum coefficient (MFCC) perception method is applied to locate speech points, and then signal processing methods are used to extract the features of the speech signal.

2 After constructing an auxiliary recognition system for spoken pronunciation, a large number of standard English spoken pronunciation audio were obtained to form a database. The original signal was decomposed using Viterbi operation to obtain the confidence value of the final intonation point. Then, the HMM classification algorithm is introduced to classify the pitch points. Due to the assumption of the HMM model that the current state only depends on the previous state, it may not be accurate enough when processing speech signals with long-range dependencies. Therefore, by selecting the method of speech segment verification, the spoken pronunciation signal is divided into multiple speech accuracy points, and combined with regularisation processing to obtain the confidence of the accuracy points, thus establishing a spoken pronunciation comparison database containing standard state sequences.

3 After adaptive filtering and clustering processing of speech key pitch point signals, a degree component signal detection method is used to obtain the spectral features of pronunciation error signals. Using the fuzzy feature state separation method to fuse the spectral features of incorrect signals and compare them with standard state sequences to capture incorrect English spoken pronunciation signals.

The innovation of this study is mainly reflected in the following aspects:

1 By comprehensively utilising speech production systems and signal processing techniques, we extract the features of speech signals and convert spoken pronunciation into processable data forms. This comprehensive application can improve the accuracy and stability of oral pronunciation analysis.

2 Using English speech recognition as the core, the confidence level of pitch points is determined through regularisation processing. This method can accurately locate and compare spoken pronunciation, providing accurate reference for subsequent pronunciation error detection.

3 In response to the shortcomings of HMM classification algorithm in processing speech signals with long-distance dependencies, a speech segment verification method is used to classify phonetic points, and a spoken pronunciation comparison database containing standard state sequences is established. Such a classification and comparison database can provide an accurate reference benchmark for comparing and capturing errors with test data.

4 Introduce a degree component signal detection method to capture pronunciation errors based on spectral features. This method can effectively identify and analyse errors in spoken pronunciation, improving the accuracy and efficiency of error capture.

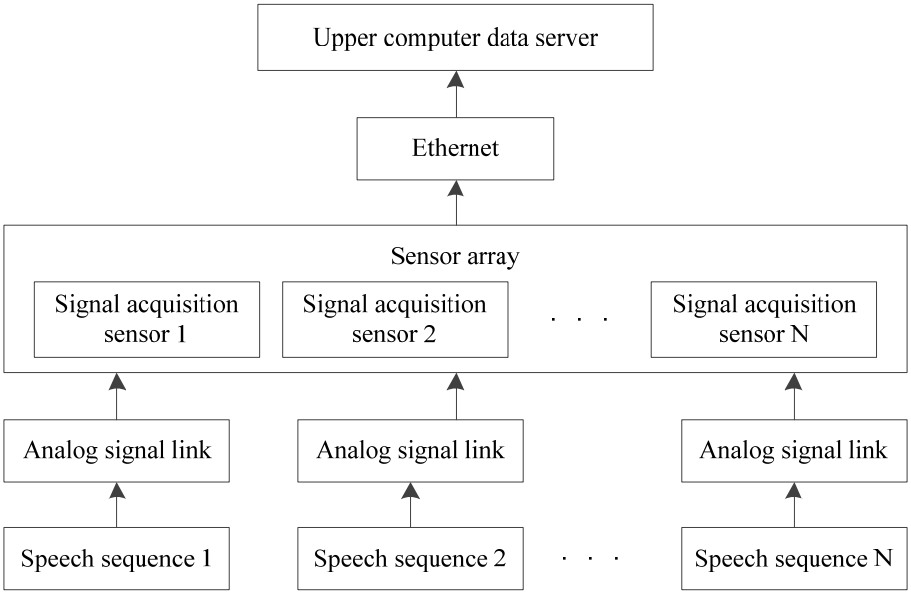
2 Preprocessing and feature extraction of English oral pronunciation signals

This study first uses a voice production system to collect raw English spoken speech data, and compares it with standardised speech signals. The English spoken speech signal to be detected is obtained through speech signal matching. Then, signal processing methods are

used to extract and recognise the features of the speech signal, and corresponding speech signal feature models are constructed.

According to the current voice signal processing requirements of the voice production system, in an Ethernet environment, the signal acquisition sensors are arranged into a uniformly distributed array, and the acquisition structure shown in Figure 1 is used to complete the acquisition of English spoken pronunciation signals.

Figure 1 Structure of English oral pronunciation signal collection



The expected goal of collecting raw English pronunciation signals is to accurately collect pure pronunciation signals. However, in practical environments, background noise may interfere with the acquisition of speech signals, affecting the quality of the signals and the accuracy of subsequent processing. For this reason, in the process of collecting English spoken pronunciation signals, in order to improve the accuracy of the collection results, this study chose a relatively quiet environment to avoid interference from background noise on the signal. After collecting the pronunciation signal, the signal was matched and filtered using the signal filtering detection transfer function to ensure excellent signal quality.

After collecting the English spoken pronunciation signal to be detected, set the signal feature detection frequency to $a(t,\alpha)$, and there are:

$$a(t,\alpha)=\sum_{N=1}^{N}\delta_i^*(\alpha)b_i(t)=\sum_{N=1}^{N}b_i^*(t)\,\delta_i(\alpha) \tag{1}$$

where * represents a complex conjugate operator; $\delta_i^*(\alpha)$ represents the calculated component of signal acquisition; $\delta_i(\alpha)$ represents the original calculation amount of signal acquisition results; $b_i(t)$ represents the component of spoken pronunciation signal. Then, adaptive algorithms are used to control the signal waveform and filter it to obtain the frequency domain characteristics of the output signal, as follows:

$$a(t,\alpha) = \delta^H(\alpha)b(t) = b^H(\alpha)\delta(t) \tag{2}$$

where H represents complex conjugate transposition; $b(t)$ represents the instantaneous signal component of English speech output; $\delta(t)$ represents the instantaneous signal weighting vector of English speech output. These two vectors can be represented as:

$$b(t) = \left[b_1(t), b_2(t), \ldots, b_n(t)\right]^T \tag{3}$$

$$\delta(\alpha) = \left[\delta_1(\alpha), \delta_2(\alpha), \ldots, \delta_n(\alpha)\right]^T \tag{4}$$

Sort out the above formulas, set the acquisition delay of English pronunciation signal to $t_0 = \sin\alpha\dfrac{\dot{r}}{r}$ in combination with relevant voice detection methods, and perform spectrum separation processing on the signal to obtain the instantaneous spectrum density ρ of English speech signal, and then establish an English speech signal model according to the spectrum. The process is as follows:

$$M = a(t,\alpha)\times\rho\times(t+t_0) \tag{5}$$

where t represents the signal processing time. On the basis of the first signal processing, in order to further remove noise from the signal, the bandwidth parameter of the high pass filter is set to $\eta_1(c)$.

For this English speech signal model, the input information is raw English spoken speech data, and the output information is speech signal features.

Because the setting result is affected by the strength of the voice signal, it is necessary to ensure that the complex conjugate coefficient is always the minimum value in the detection process. Based on this setting, a signal filtering detection transfer function γ is introduced, which is used to complete signal matching and filtering processing, eliminate negative frequencies in the signal, convert the real signal into a complex signal, and obtain the processed speech pronunciation signal as follows:

$$V = \frac{M\times\eta_1(c)}{\gamma} \tag{6}$$

Then, the MFCC perception method (Bai et al., 2021; Oh et al., 2021; Barhoush et al., 2023) is applied to locate speech points. Based on the obtained signal model, set its spectral feature to $O = \{o_{k1}, o_{k2}, \ldots, o_{ki}, \ldots, o_{kn}\}$, and apply wavelet transform to convolution O, resulting in:

$$Q_i = \{q_{k1}, q_{k2}, \ldots, q_{ki}, \ldots, q_{kn}\} \tag{7}$$

The process in formula (7) can adjust the pronunciation signal to a monosyllabic signal, and apply the preset smoothing coefficient to process the signal at $\frac{N}{(N+1)}$ sampling points. The recognition results of English spoken intonation point features can be expressed as:

$$C(t) = Q_i\times\left(1-\cos(2\pi t)\right) + V\times\left(1+T\cos(2\pi t+\varphi)\right) \tag{8}$$

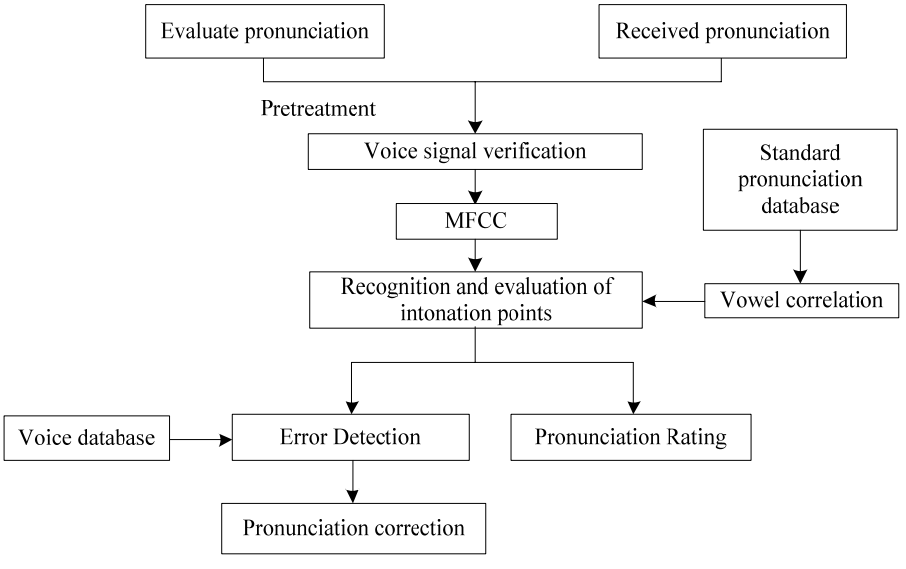
Among them, φ represents the smoothing coefficient; T represents the signal processing cycle. Apply the above formula to obtain the smoothed pronunciation signal features and use them as the basis for subsequent processing. After obtaining the signal model, it is necessary to digitise it and convert it into a digital signal for later analysis and capture.

The above process utilises a speech production system to collect raw speech data, and obtains the English spoken pronunciation signal to be detected through speech signal matching, thereby completing signal feature recognition. Next, based on the feature recognition results, we will construct an auxiliary recognition system for oral pronunciation and a comparison database for oral pronunciation, laying the foundation for capturing pronunciation errors in the future.

3 Building an English speech recognition system

To better capture errors in English spoken pronunciation, a system for auxiliary recognition of spoken pronunciation is constructed based on the recognition results of English spoken intonation point features, with the basic structure shown in Figure 2.

Figure 2 Basic structure of English speech recognition system



In Figure 2, the system takes speech segment recognition and pronunciation evaluation as the core links, and in this chapter, we will mainly design these two parts.

Based on the preset basic framework, first obtain a large number of correct English spoken pronunciation audio, establish a database, and provide a foundation for subsequent pronunciation error detection and capture. In this study, regularisation processing is used to determine the confidence level of intonation points, and then HMM classification algorithm (Jin et al., 2021; Ali et al., 2021; Na and Park, 2021) is used to classify intonation points, so as to establish a database of spoken English pronunciation comparison.

The steps to design a voice comparison database containing standard state sequences are as follows:

Step 1: Collect standard pronunciation samples. Collect a series of standard English oral pronunciation samples based on the required accent or pronunciation standards. These samples can be recorded by professional pronunciation professionals to ensure accurate pronunciation and meet expected standards.

Step 2: Mark the Standard state sequence. Label the phonemes in each speech sample and identify the start and end time points of each phoneme. In this way, the Standard state sequence of each sample can be obtained, which corresponds to the text or audio.

Step 3: Extract feature vectors. For each labelled sample, the corresponding feature vectors are extracted through signal processing technology. Common features include MFCC, Mel spectrum, Zero-crossing rate, etc.

Step 4: Establish a database. The eigenvector of each sample and its corresponding Standard state sequence are stored as records in the database. In this way, a voice comparison database containing Standard state sequences is formed.

Before putting it into use, it is necessary to evaluate the accuracy of the database. This study adopts a combination of manual evaluation and validation set testing to test, evaluate, and adjust the database to ensure its accuracy and reliability. Professional reviewers manually review and rate the samples in the database. Reviewers evaluate whether the pronunciation in the database is accurate based on standard pronunciation or specific pronunciation requirements, and what is the expected accuracy rate in the design. At the same time, randomly select a portion of samples from the database as the validation set to test the recognition accuracy of the algorithm on the validation set. This can provide more objective results and also evaluate the generalisation ability of the algorithm on different sample sets.

On this basis, based on the Standard state sequence in the database, in the subsequent detection, the pronunciation frame sequence to be detected is compared with the Standard state sequence, which can achieve accurate capture of spoken English pronunciation errors.

Assuming that the spoken vocabulary in the database is *S*, the number of HMM parameters in this study is *S*. Then, use these parameters to perform Viterbi operations and determine the state sequence of each pronunciation model. Then, using these parameters, the Viterbi operation is performed to determine the state sequence of each pronunciation model as follows:

$$G_i\left(x_f\right)=\frac{\exp\left\{-\frac{1}{2}\left(x_f-x_{DR}\right)^T D_i^{-1}\left(x_f-x_{DR}\right)\right\}}{\sqrt{\left|D_i\right|(2\pi)^v}} \tag{9}$$

Among them, v represents the dimension of the feature vector; x_f represents the feature vector; x_{DR} represents the average value of the state; D_i represents the covariance matrix of the density function; $G_i\left(x_\varepsilon\right)$ represents the similarity probability between the feature vector x_ε and the state i.

However, HMM model assumes that the current state only depends on the previous state, which may not be accurate enough when dealing with speech signals with long-

range dependencies. Therefore, this study chooses the method of speech segment verification to determine the correctness of speech content based on the speech judgement threshold. Split the spoken pronunciation signal into multiple speech accuracy points, perform regularisation on these points, and obtain the confidence level of the accuracy points. After splitting the pronunciation signal into pitch points, it is easier to locate specific pronunciation errors, which is very helpful for pronunciation training and correction. The engineering settings for regularisation processing are as follows:

$$X_{\sigma}=\frac{2}{\left(u\times T\times\left(\frac{\lg_n^w E}{\lg_n^w E_0}\right)\right)_{1+e}} \tag{10}$$

Among them, $\frac{\lg_n^w E}{\lg_n^w E_0}$ represents the probability difference of vowel logarithms; X_{σ} represents the confidence calculation result; u represents adjusting the parameter value. By using this formula, the confidence level of each pitch can be calculated. Due to the similarity of most syllables in spoken English (Lee et al., 2021; Huang et al., 2021), when calculating confidence, it is not only necessary to rank the confidence, but also to analyse the difference in pitch points to obtain the final pitch confidence value. The specific calculation formula is set as follows:

$$X_{\sigma}^{*}=X_{\sigma}\times 100\times\sum_{1}^{n}\frac{T\times(Vowel)}{\left(u\times T\times\left(\frac{\lg_n^w E}{\lg_n^w E_0}\right)\right)} \tag{11}$$

Among them, n represents the number of intonation points.

Organise the above settings and build a database of English spoken pronunciation comparison. In the process of speech recognition, the probability of each pronunciation model is determined based on the preset parameters mentioned above, and the optimal state sequence is found. When this frame sequence corresponds to the state sequence, it can effectively avoid time distortion in the current speech signal processing process and improve the capture effect (Savchenko and Savchenko, 2022).

The above process classifies the intonation points by using the HMM classification algorithm, and establishes a spoken pronunciation comparison database containing Standard state sequences, which helps to establish a comparison reference, so as to better judge whether the pronunciation is correct.

4 Comparison and capture of English oral pronunciation errors

Based on the above designed English speech recognition system, the original spoken speech signals are organised into two parts: pronunciation rhythm and basic speech. The acoustic features of pronunciation speech bars are extracted, and they are recognised according to a preset resource library to obtain the final detection phoneme model. Through comparison, the capture of incorrect pronunciation is achieved. The technical depth of this link is reflected in the following aspects:

1 By performing adaptive filtering and focusing on key pitch point signals in speech, spectral features related to pronunciation errors can be effectively extracted and highlighted. This processing method uses adaptive filters to enhance or suppress signals within a specific frequency range, thereby making the characteristics of erroneous signals more significant and clear.

2 Effectively analyse and capture the spectral features of erroneous signals using degree component signal detection methods. By comparing the spectral features of incorrect signals with standard state sequences, it is possible to identify and capture pronunciation error signals.

3 By comparing the spectral features of incorrect signals with standard state sequences and using fuzzy logic for feature state separation, it is possible to more accurately distinguish and capture pronunciation error signals. The application of this method reflects the depth and complexity of technology.

After studying a large number of literature, it was found that there are two types of current pronunciation detection methods: one requires the use of linguistics to capture errors using new language features; Another way to capture errors is through speech recognition, which involves comparing databases and calculating confidence levels (Wu et al., 2022; Tabet, 2022; Kaur et al., 2022). With the development of speech processing technology, this study highly integrates the existing speech recognition technology and proposes a more suitable capture method for English spoken language.

After completing the filtering process, the speech signal is reintroduced into the speech recognition system. After removing interference noise, the key pitch point signal of the speech is adaptively filtered and focused, which includes:

$$l_i(t) = y_i(t)\cos\left[2\pi f^i t + \varphi_i(t)\right] \tag{12}$$

Among them, $y_i(t)$ represents the complex signal of the forward frequency part of the spectrum; $\varphi_i(t)$ represents the noise subspace of the forward frequency part of the spectrum; f^i represents the noise function. Select the beam adaptive focusing method to obtain the transfer function of the speech pitch recognition point signal:

$$Z(s) = \frac{\sin\gamma_2 + \sin\gamma_1\left(1+\sin\gamma_2\right)s^{-1} + s^{-2}}{1+\sin\gamma_1\left(1+\sin\gamma_2\right)+\sin\gamma_2 s^{-2}} \tag{13}$$

Among them, γ_1 represents the frequency spectrum of the initial signal of English spoken syllables; γ_2 represents the frequency spectrum of syllable tail signals in spoken English; s^{-1} represents the value of signal transmission parameters; s^{-2} represents the value of signal transmission parameters after adaptive focusing processing. For all speech signals, the pronunciation spectrum exists $\left|Z(m)=1\right|$, and the spectral components of known pronunciation errors should meet $Z(m)=m$. The degree component signal detection method is used to obtain the spectral characteristics of the output oral pronunciation error signal:

$$\widetilde{l_i(t)} = y_i(t)exp[j(2\pi f^i t + \varphi_i(t))] \tag{14}$$

where $y_i(t)$ represents the spectral component; $\varphi_i(t)$ represents the preset error signal spectrum parameters. After obtaining spectral features, scale decomposition and feature extraction are performed on the collected spoken speech signals, and pronunciation errors are captured and detected based on this. To avoid overfitting, a fuzzy feature state separation method (Savchenko, 2022; Savchenko and Savchenko, 2021) is used to obtain the error speech feature parameters $h_1(t)$ and $h_2(t)$, which can be expressed as follows:

$$\begin{cases} h_1(t) = -2x'(t)\cos\left(\tau(t)\right) \\ h_2(t) = x'^2(t) \end{cases} \tag{15}$$

After fusing it with spectral features, the final pronunciation error feature screening result is obtained:

$$\begin{cases} h(t) = \sqrt{l'^2(t) + b'(t)} \\ \mu(t) = \arctan\left\{\dfrac{b'(t)}{l'(t)}\right\} \end{cases} \tag{16}$$

Among them, $h(t)$ represents the instantaneous amplitude of the spoken pronunciation pitch point signal; $\mu(t)$ represents the fuzzy component of English spoken pronunciation errors. Use the following formula to obtain the pronunciation error feature threshold:

$$\begin{cases} \beta'_{\min,j} = \max\left\{\beta'_{\min,j}, \beta'_{h,j} - \tau\left(\beta'_{\max,j} - \beta'_{\min,j}\right)\right\} \\ \beta'_{\max,j} = \min\left\{\beta'_{\max,j}, \beta'_{h,j} - \tau\left(\beta'_{\max,j} - \beta'_{\min,j}\right)\right\} \end{cases} \tag{17}$$

Among them, $\beta'_{\min,j}$ represents the minimum eigenvalue of incorrect pronunciation; $\beta'_{\max,j}$ represents the maximum eigenvalue of incorrect pronunciation; τ represents the spectral characteristics of spoken voice error signals. Combined with prior probability, the final result of capturing errors in spoken English pronunciation is obtained:

$$w'_i(p', \sigma') = w'_h(p', \sigma') + \varepsilon \sum_{i=0}^{\infty} z'(i) z'^3(i + \sigma') \tag{18}$$

Among them, σ' represents the component of English spoken pronunciation errors; $w'_i(p', \sigma')$ represents the characteristic value of English spoken pronunciation errors.

Integrate the above content and introduce the speech recognition detection part into the English speech recognition system, constructing a complete speech recognition and pronunciation error capture process.

5 Experiment and result analysis

To test the feasibility of a speech recognition based method for capturing English oral pronunciation errors, an experimental section is constructed to analyse the capture ability and effectiveness of this method.

5.1 Experimental plan design

This study completed the specific operation and analysis process through simulation experiments, using Matlab7 simulation software as the experimental foundation. During this experiment, the number of sampling nodes for English spoken pronunciation signals was set to 100, the feature extraction resolution for pronunciation signals was set to 150 KHz, the length of the output English spoken signal was 1000, the interference signal-to-noise ratio was set to 10dB, the filter order was set to 10, and the average frequency of the signal spectrum was set to 1.50 KHz. Based on this setting result, set up the experimental phase. In addition to the above content, the experimental environment is set up and simulated according to the daily English oral teaching environment and English oral testing environment to ensure the effectiveness of the experimental results.

5.2 Experimental data

This experiment takes a certain university as an example, selecting some non-native language learners as experimental volunteers and providing them with an English oral corpus for text reading. This part of the speech is obtained as the data basis for subsequent experiments. At the same time, the correct pronunciation signals from the spoken language corpus are added to the speech database as the database of English speaking testers for this experiment. To reduce the difficulty of experimental operations, the collected audio was divided into 5 groups, each containing incorrect and correct pronunciation signals. The specific experimental signal group division results are shown in Table 1.

Table 1 Experimental signal group division results

Experimental group serial number	*Vocabulary/ piece*	*Short sentences/ piece*	*Total/piece*	*Number of incorrect pronunciation/ piece*	*Number of correct pronunciation/ piece*
1	395	105	500	150	350
2	505	120	625	220	405
3	700	84	784	154	630
4	510	400	910	170	740
5	200	162	362	105	257

Based on the data in Table 1, analyse and evaluate the application effectiveness of the capture method proposed in this paper. To avoid the singularity of experimental results, the method of Huo and Xie (2022) and method of Jin et al. (2021) are compared and analysed through comparative analysis of experimental results to determine whether the method of this paper has achieved the design objectives.

5.3 Experimental indicator setting

In this experiment, the experimental indicators were set as the accuracy of capturing oral pronunciation errors, the accuracy of recognising oral pronunciation signals, and the

capture time of oral pronunciation errors as evaluation criteria for the effectiveness of different methods. The specific experimental index calculation process is set as follows:

1 *Recognition accuracy of spoken pronunciation signals*

Due to the long and complex pronunciation signals of some spoken English, some current methods cannot effectively recognise long and difficult sentences or short sentences, resulting in the inability to capture high-quality speech samples. Therefore, the recognition accuracy of spoken English pronunciation signals is used as an experimental indicator, and the specific formula is set as follows:

$$\check{p} = 100\% \times \frac{Q'_I}{Q'} \tag{19}$$

Among them, Q'_I represents the number of recognised pronunciation signals; Q' represents the number of preset pronunciation signals.

2 *Accuracy of capturing oral pronunciation errors*

$$\check{J} = 100\% \times \frac{M'_I}{M'} \tag{20}$$

Among them, M'_I represents the number of captured incorrect pronunciation signals; M' represents the preset number of incorrect pronunciation signals.

In this experiment, the initial value of the number of iterations for capturing pronunciation errors was set to 20. Each experiment was conducted with an additional 20 iterations until the number of iterations increased to 60. The accuracy of capturing oral pronunciation errors for different methods under each iteration condition was calculated to determine the capture effect and computational ability of different methods.

3 *Average time to capture oral pronunciation errors*

$$V' = \frac{T'_1 + T'_2 + \cdots + T'_N}{N'} \tag{21}$$

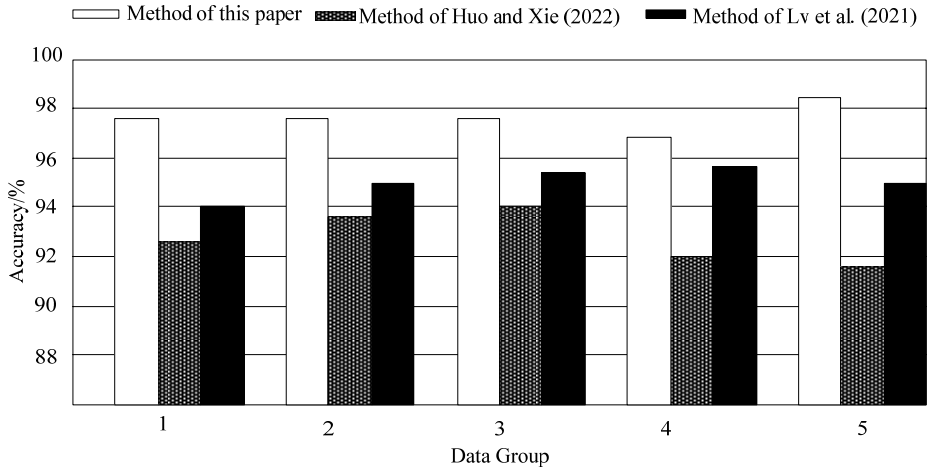
Among them, T' represents the capture time of a single experimental pronunciation error signal; N' represents the number of pronunciation signals to be detected. For the convenience of subsequent processing, obtain the pronunciation error capture time of each group under different iterations, and calculate the average value as the final experimental result output.

5.4 Test results and analysis

5.4.1 Analysis of experimental results on the accuracy of oral pronunciation signal recognition

Firstly, analyse the recognition accuracy of different methods for English spoken pronunciation signals, and the results are shown in Figure 3.

Figure 3 Experimental results on the accuracy of oral pronunciation signal recognition



Analysing Figure 3, it can be seen that during the preliminary process of recognising spoken pronunciation signals, the method of this paper can fully recognise the speech signals in the experimental set, and the recognition accuracy remains above 97%. The recognition accuracy of Lv et al. (2021) for English spoken pronunciation signals remains above 94%, slightly lower than the method of this paper. The method of Huo and Xie (2022) has the worst recognition ability for speech signals. The reason for the high recognition accuracy of this method is that it is based on the English speech recognition system and determines the confidence level of pitch points through regularisation processing. This can effectively help locate key pronunciation points in speech, thereby more accurately capturing pronunciation errors and improving recognition accuracy.

5.4.2 Analysis of experimental results on the accuracy of capturing oral pronunciation errors

Then, analyse the accuracy of capturing English oral pronunciation errors using different methods, and the results are shown in Table 2.

Analysing the data in Table 2, it can be seen that as the number of iterations continues to increase, the accuracy of capturing oral pronunciation errors by different methods has been correspondingly improved. However, through horizontal comparison, it can be seen that the method of this paper can achieve high capture accuracy in both low iteration environment and high iteration environment, with a maximum accuracy of 98.74%. This experimental result indicates that the method of this paper will not be affected by the external environment and the computing power of the detection system during the application process. Compared with the method of this paper, the capture accuracy of the other two methods is much lower than that of the method of this paper, and the accuracy fluctuates greatly. Based on the above experimental results, it can be determined that the method of this paper has a relatively high ability to capture pronunciation errors. This is because this method introduces a degree component signal detection method, which can accurately capture English spoken pronunciation error signals by comparing spectral features with standard state sequences. In addition, the method proposed in this paper innovatively uses the HMM classification algorithm to classify intonation points based on

the establishment of a spoken pronunciation comparison database. This method can further improve accuracy by comparing with standard state sequences.

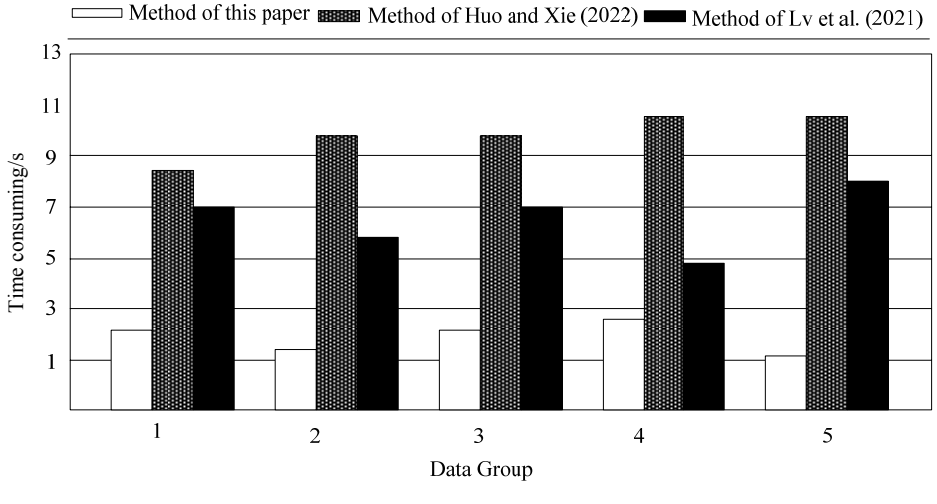
Table 2 Experimental results on the accuracy of capturing oral pronunciation errors (Unit:%)

Number of iterations/time	*Experimental group serial number*	*Method of this paper*	*Method of Huo and Xie (2022)*	*Method of Lv et al. (2021)*
20	1	95.15	90.32	93.15
	2	95.87	91.14	93.44
	3	96.00	91.65	94.56
	4	95.18	92.65	94.68
	5	95.27	93.15	95.31
40	1	96.15	94.30	96.01
	2	96.67	94.65	97.14
	3	96.74	93.25	97.65
	4	97.85	91.54	96.88
	5	98.15	93.56	96.14
60	1	98.15	97.18	97.52
	2	98.22	97.22	97.14
	3	98.37	97.64	97.68
	4	98.64	96.18	97.81
	5	98.74	97.85	97.22

5.4.3 Analysis of experimental results on the average time for capturing oral pronunciation errors

Finally, the average capture time of oral pronunciation errors for different methods was analysed to verify their timeliness, as shown in Figure 4.

Figure 4 Experimental results of average time to capture oral pronunciation errors



Analysing Figure 4, it can be seen that there are significant differences in the timeliness of the three methods in the application process. Method of Lv et al. (2021) requires multiple filtering processes on the original signal and multiple compensations when capturing pronunciation signals, resulting in a longer time for capturing pronunciation errors. Method of Huo and Xie (2022) overly relies on the computing power of the detection system. If the detection system has poor computing power, it will cause serious pronunciation error capture delay issues. And this method can control the capture time within 3 s, indicating that this method can more quickly capture oral pronunciation errors. This is because based on the construction of a speech recognition system, the method in this paper has classified and processed the intonation points, established a comparison database, and processed the speech signals quickly through comparison, so that English oral pronunciation errors can be captured in the shortest possible time.

6 Conclusion

Applying speech recognition technology to the process of capturing English oral pronunciation errors can effectively improve the reliability of the results of capturing English oral pronunciation errors. Therefore, based on the construction of a speech recognition system, this study applies this system as the core and foundation for capturing English oral pronunciation errors. Through speech signal processing, speech recognition, speech signal feature collection and comparison, English oral pronunciation error detection is achieved.

The experimental results show that the method of this paper has further improved the accuracy of pronunciation error capture results and shortened error capture time after applying speech recognition technology. The experimental results confirm the scientific and effective nature of the proposed method in this study.

Although this method performed well in experiments, its performance may be affected by different speech characteristics. For example, factors such as different people's accents, speech rates, volume changes, and emotional expressions can all lead to a decrease in recognition accuracy. In addition, the pronunciation characteristics of children or elderly people may differ from those of adults, further increasing the difficulty of recognition. Therefore, in the following research, the adaptability of the model to speech diversity can be improved by expanding the diversity of the training dataset, including speech samples of different genders, ages, accents, and emotional expressions.

References

Ali, A., Chowdhury, S., Afify, M., El, H.W., Hajj, H., Abbas, M., Hussein, A., Ghneim, N., Abushariah, M. and Alqudah, A. (2021) 'Connecting Arabs: bridging the gap in dialectal speech recognition', *Communications of the ACM*, Vol. 64, No. 4, pp.124–129.

Bai, S., Yan, X.H., Zhang, S.P., Chne, J.F. and Zhang, S. (2021) 'Voice activity detection based on Mel frequency cepstrum coefficient and short time energy in low SNR', *Journal of Nanjing Normal University (Natural Science Edition)*, Vol. 44, No. 2, pp.117–120.

Barhoush, M., Hallawa, A. and Schmeink, A. (2023) 'Speaker identification and localization using shuffled MFCC features and deep learning', *International Journal of Speech Technology*, Vol. 26, No. 1, pp.185–196.

Huang, Y., Zhao, F.H. and Lu, J. (2021) 'An improved algorithm of double threshold endpoint detection method in speech signal processing', *Acta Scientiarum Naturalium Universitatis Nankaiensis* (*Natural Science Edition)*, Vol. 54, No. 2, pp.58–62.

Huo, M.H. and Xie, Y.L. (2022) 'Speech attributes optimization and its application in mispronunciation detection', *Journal of Chinese Information Processing*, Vol. 36, No. 1, pp.163–172.

Jin, Y., Li, Y.G. and Ji, H.B. (2021) 'Adaptive ASR filtering in impulsive noise environments', *Journal of Electronics and Information Technology*, Vol. 43, No. 2, pp.296–302.

Kaur, I., Nassa, V.K., Kavitha, T., Mohan, P. and Velmurugan, S. (2022) 'Maximum likelihood based estimation with quasi oppositional chemical reaction optimization algorithm for speech signal enhancement', *International Journal of Information Technology*, Vol. 14, No. 6, pp.3265–3275.

Lee, D., Kim, D.H., Yun, S. and Kim, S.H. (2021) 'Phonetic variation modeling and a language model adaptation for Korean English code-switching speech recognition', *Applied Sciences*, Vol. 11, No. 6, pp.2866–2866.

Lei, J., Zhao, H.L., Ai, N.Z., Zou, W.B. and Zhan, Y. (2021) 'Low-resource speech recognition system based on BN-SGMM-HMM model', *Journal of Hefei University of Technology (Natural Science)*, Vol. 44, No. 12, pp.1627–1632.

Lv, S.R., Wu, C.G., Liang, Y.C., Yuan, Y.P., Ren, Z.M., Zhou, Y. and Shi, X.H. (2021) 'An end-to-end Chinese speech recognition algorithm integrating language model', *Acta Electronica Sinica*, Vol. 49, No. 11, pp.2177–2185.

Na, H.J. and Park, J.S. (2021) 'Accented speech recognition based on end-to-end domain adversarial training of neural networks', *Applied Sciences*, Vol. 11, No. 18, pp.8412–8412.

Oh, D.H., Park, J.S., Kim, J.W. and Jang, G.J. (2021) 'Hierarchical phoneme classification for improved speech recognition', *Applied Sciences*, Vol. 11, No. 1, pp.428–428.

Savchenko, A.V. and Savchenko, V.V. (2022) 'Adaptive method for measuring a fundamental tone frequency using a two-level autoregressive model of speech signals', *Measurement Techniques*, Vol. 65, No. 6, pp.453–460.

Savchenko, V.V. (2022) 'Method for reduction of speech signal autoregression model for speech transmission systems on low-speed communication channels', *Radioelectronics and Communications Systems*, Vol. 64, No. 11, pp.592–603.

Savchenko, V.V. and Savchenko, L.V. (2021) 'Speech signal autoregression modeling based on the discrete Fourier transform and scale-invariant measure of information discrimination', *Journal of Communications Technology and Electronics*, Vol. 66, No. 11, pp.1266–1273.

Tabet, Y. (2022) 'Speech signal analysis with a refined iterative adaptive method', *International Journal of Electronics, Communications, and Measurement Engineering* (*IJECME*), Vol. 11, No. 1, pp.1–18.

Wang, H.Y., Jeon, E., Zhang, W.Q., Li, K. and Huang, Y.K. (2023) 'Zero resource Korean ASR based on acoustic model sharing', *Journal of Data Acquisition and Processing*, Vol. 38, No. 1, pp.93–100.

Wang, P., Bai, J. and Xue, P.Y. (2021) 'Pathological speech detection with articulatory movement features and acoustic features', *Computer Engineering and Design*, Vol. 42, No. 3, pp.776–781.

Wu, J., Chen, X.L., Shi, S.J., Guo, X.B. and Fan, L. (2022) 'Application of adaptive low-pass filtering method to reduce noise signal in radiation measurement', *Nuclear Electronics and Detection Technology*, Vol. 42, No. 1, pp.155–160.

Yu, Y.D., Li, C.J. and Yang, L. (2022) 'Intelligent home internet of things system based on speech recognition', *Journal of Computer Applications*, Vol. 42, No. S1, pp.391–394.