Automatic Pronunciation Calibration Method of Language Resource Base Based on Dynamic Time Rounding Algorithm

Shao Gong
School of Humanities and Communication
University of Sanya, China
lll6544511@yeah.net

Heng Xiao
School of Information and Intelligence Engineering
University of Sanya, China
xiaoheng989@163.com

Abstract: *The pronunciation accuracy of language resource library is the key to improve the quality of language resource library. An automatic pronunciation calibration model construction method based on dynamic time normalization algorithm is proposed. By analyzing the dynamic characteristics of the pronunciation of the language resource library, the acquisition model of the pronunciation of the language resource library is constructed, and the pronunciation of the language resource library is obtained. The ambiguity detection method is used to suppress the noise of the pronunciation signal of the language resource library. According to the processing results, the speech interaction method of the language resource library is used to analyze the matching domain of the voice signal obtained by interval uniform sampling, and extract the voice characteristics of the language resource library, Based on the extracted features, the dynamic time normalization algorithm is used to recognize the speech similarity, and the voice signal detection model of the optimal language resource library is established under the given false alarm probability, so as to improve the automatic voice calibration capability of the language resource library within the prior Doppler frequency range. The simulation results show that this method has a high accuracy probability and a low false alarm probability for speech detection in the language resource library, which improves the ability of speech interaction and dynamic feature analysis of the language resource library*.

Keywords: *Dynamic time rounding, voice interaction, semantics, language resource library, pronunciation, detection model*.

Received February 20, 2025; accepted August 5, 2025
https://doi.org/10.34028/iajit/23/1/15

1. Introduction

In language teaching, language resource pool is of great significance [2, 13, 16]. Through long-term reading training of language resource pool, vocabulary accumulation and pronunciation accuracy of language can be improved, and reading ability can be further improved. In the design of language resource pool, the pronunciation standard of language resource pool is the key. Combined with the feature detection and recognition of pronunciation signal parameters of language resource pool [17, 18] multi-dimensional feature classification detection is adopted to realize pronunciation signal recognition of language resource pool, and the ability of pronunciation signal detection and dynamic parameter analysis of language resource pool is improved, so as to improve the accurate detection ability of pronunciation signal of language resource pool. It is of great significance to study the optimal construction method of automatic pronunciation calibration model of language resource pool to improve the application function of language resource pool.

Automatic pronunciation calibration of language resource pool is based on the analysis of pronunciation signal characteristics and information parameter identification of language resource pool. In traditional methods, the automatic pronunciation calibration methods of language resource pool mainly include automatic pronunciation calibration method of language resource pool based on envelope statistical information fusion [7, 9, 15], automatic pronunciation calibration method of language resource pool based on phase space reconstruction, etc. Signal recognition and signal feature parameter recombination methods are adopted to realize the pronunciation signal feature fusion of language resource pool.

Dong *et al*. [6] proposes a speech enhancement algorithm based on waveform mapping in time domain and harmonic loss in frequency domain. The Harmonic Noise Model is used to model the pure speech, and the HNM component in the frequency domain obtained after modeling is used as the training target in the loss function. By minimizing the harmonic loss function in frequency domain, the full convolution neural network is trained to produce time domain enhanced speech. However, this method trains a fully convolutional network through time-domain waveform mapping and frequency-domain harmonic loss, but does not explicitly compensate for speech phase. Due to the sensitivity of the human ear to phase, phase distortion can lead to a decrease in the naturalness of speech, especially at low signal-to-noise ratios. The phase interference of noisy speech can significantly reduce the calibration effect. A real-time speech enhancement algorithm that can resist unsteady noise is proposed by Xiao and Chen [20]. The band gain estimation in RNNoise is converted into the prior signal-to-noise ratio of the band as the input feature of neural network, and the harmonic gain is corrected with the pitch detection algorithm to reduce the deviation of gain estimation. However, algorithms rely on pitch detection algorithms to correct harmonic gains, but in strong noise environments, pitch period estimation may be inaccurate, leading to gain correction bias and causing speech distortion or noise residue. Wang *et al*. [19] proposed first adopted the traditional dual window size method to achieve frame online speech enhancement. Next, complex spectrum mapping will be used for frame online enhancement, where Deep Neural Networks (DNN) learns how to extract spectral information of target speech from mixed speech signals. Then, the RI component predicted by DNN is used for frame online beamforming. The results of beamforming are used as additional features for the second DNN to perform frame online post filtering. Post filtering is aimed at further improving the quality of speech signals. The technique of extracting and enhancing target speech from mixed speech signals. However, when the beamforming result is used as the input for post filtering, if the optimization objectives of the two are inconsistent (such as beamforming focusing on directionality and post filtering focusing on residual noise suppression), it may introduce signal distortion or excessive smoothing, affecting speech clarity calibration. Lin *et al*. [11] proposed an adaptive noise distribution network speech enhancement algorithm. An Adaptive Gaussian Unitary Ensemble Attention (SA-GUEA) block was constructed in the SASE network, enabling the model to handle noise more intelligently. Develop optimization weighting strategies based on loss and Perceptual Evaluation of Speech Quality (PESQ). The server model can be intelligently updated to better generalize when dealing with large-scale heterogeneous datasets. However, although SA-GUEA blocks can improve the intelligence of noise processing, complex attention mechanisms may increase computational latency, leading to a decrease in real-time calibration performance, especially when dealing with high-speed time-varying noise, where calibration lag may affect speech coherence. Zhang [22] designed a system hardware using speech perception sensors and English oral pronunciation processors to collect digital signals of English oral pronunciation. Based on this, the signals were pre emphasized, framed, and windowed to obtain signal feature MFCC coefficients. The MFCC coefficients were used as the training dataset to construct a pronunciation error detection model, which automatically calibrated English oral pronunciation based on detected erroneous pronunciations. However, the pronunciation error detection model is trained based on MFCC coefficients, and its accuracy directly determines the quality of calibration performance. If there is insufficient training data, improper feature selection, complex model structure, or non-convergence of training algorithms, the model may not be able to accurately detect pronunciation errors, resulting in a decrease in calibration performance. Zheng *et al*. [23] designed a virtual reality based English pronunciation calibration simulation system. The system consists of client and server-side modules. The client module provides users with an interactive interface and obtains user control commands, while the server-side module collects user commands, responds, and effectively processes virtual simulation scene business; The system achieves comprehensive and accurate calibration of English pronunciation through the English pronunciation calibration process, and completes precise calibration of English pronunciation through error correction calculations. However, virtual reality systems typically involve complex graphics rendering and interactive processing, which may result in system latency. If the system latency is too high, users may experience significant lag or asynchrony during the pronunciation calibration process, which can affect their experience and calibration effectiveness.

In order to solve the above problems, this paper proposes a method to build the pronunciation automatic calibration model of language resource pool under the semantic corpus fusion scheduling environment based on dynamic time rounding algorithm. The structure of this article is as follows:

1. The construction of the pronunciation acquisition model for the language resource library was elaborated in detail, including the hardware circuit design for dynamic signal acquisition and software level spectrum parameter analysis. And through fractional spectrum analysis and phase rotation component modeling, the preprocessing of pronunciation signals was carried out, laying the foundation for subsequent feature extraction.
2. Based on the preprocessed speech signal mentioned above, speech feature extraction was achieved through time-frequency analysis and segmented frequency modulation signal decomposition.

The Dynamic Time Warping (DTW) algorithm is used to achieve speech similarity recognition, and a pronunciation detection model under semantic corpus fusion scheduling is proposed. This model combines autocorrelation analysis and noise spectrum separation to achieve accurate deviation estimation between actual pronunciation and standard templates.

Finally, the effectiveness of the proposed method was validated through waveform comparison, bias estimation visualization, and error quantification using the NOISEX dataset and controlled signal-to-noise ratio environment. The comparative experiments with existing algorithms have demonstrated the superiority of the DTW based method in terms of calibration accuracy and robustness.

2. Collection Model and Pretreatment of Language Pronunciation in Resource Bank

2.1. Pronunciation Collection Model of Language Resource Pool

To construct an automatic pronunciation calibration model of language resource pool, firstly, the dynamic characteristics of language resource pool pronunciation are analyzed, and the signal detection and spectrum parameter analysis model of language resource pool pronunciation is established by adopting linear frequency modulation signal detection technology. Through fractional spectrum analysis and the method of combining acoustic sensors, analyze the frequency modulation parameters of the pronunciation output of the language resource pool under the semantic corpus fusion scheduling environment, and realize the gradient weighted class activation of the speech signal by constructing the speech receiver shown in Figure 1. In Figure 1, $U_s(t)$ is the input voltage of the acquisition terminal, and V_g is the pronunciation vibration signal of the language resource pool under the semantic corpus fusion scheduling environment. U_4 is the differential pressure sensing variable, V_d is the dynamic output voltage of language resource pool pronunciation, Z_{in} is the input voltage, Z_L is the pronunciation control inductor of language resource pool, and Z_d is the hardware reset output inductor. The acquisition of pronunciation signals in the language resource library was achieved dynamically in the sensor power supply circuit and amplifier circuit module through the fusion scheduling environment of later circuit semantic corpus.

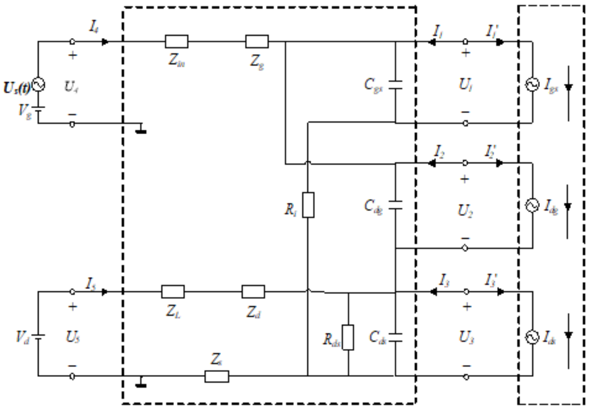
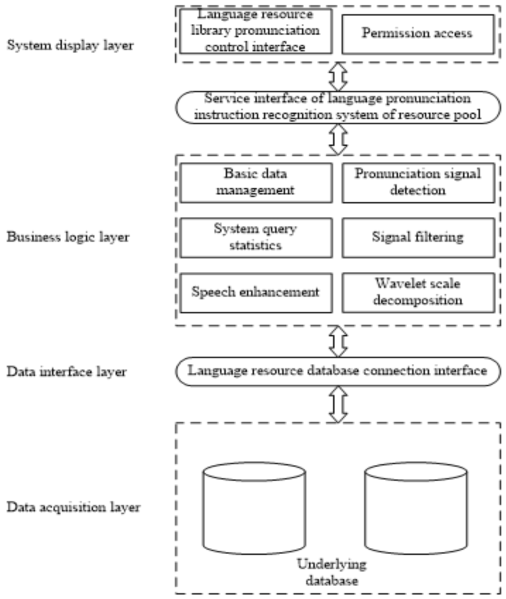


Figure 1. Dynamic acquisition circuit of pronunciation signal of language resource pool.

Figure 1 combined with the circuit structure of dynamic acquisition of pronunciation signals of language resource pool, the sensor network middleware and platform software are designed by means of sensor power supply and signal conversion, and the network access service and network generation service of pronunciation of language resource pool are completed through signal feature analysis, and the pronunciation signal model of language resource pool is established. Autocorrelation detector and matched filter detection method are used to detect and analyze the pronunciation signal characteristics of language resource pool [14, 23] and the pronunciation signal accuracy detection model of language resource pool is divided into four layers, namely, system display layer, business logic layer, data interface layer and voice data acquisition layer. The function of the display layer is to interact with users, show them information and receive their input. For the integration between systems, the role of the presentation layer is to interface with other systems. Therefore, sometimes the presentation layer is also called the service access layer, which is responsible for the docking service with other systems. The business logic layer is the core value part of the system architecture. It is located between the data access layer and the presentation layer, and plays a connecting role in data exchange. It is responsible for defining business logic rules, workflow, data integrity, etc., receiving data requests from the presentation layer, submitting requests to the data access layer after logical judgment, and delivering data access results. The data interface layer can be used to access various databases. Through this interface layer, unified instructions can be sent to the general interface, and then the instructions can be transmitted to any type of database by the interface layer, which defines a lightweight and consistent condition for accessing databases. The voice acquisition layer can provide transmission media and connections for data communication between terminal devices based on the link. In the process of collecting voice data, the transmitter and receiver can communicate with each other for one or more times, which improves the ability of data error detection and correction. Thus, the overall structure of pronunciation signal accuracy detection of language resource pool is shown in Figure 2.



System display layer
Language resource library pronunciation control interface
Permission access
Service interface of language pronunciation instruction recognition system of resource pool
Business logic layer
Basic data management
Pronunciation signal detection
System query statistics
Signal filtering
Speech enhancement
Wavelet scale decomposition
Data interface layer
Language resource database connection interface
Data acquisition layer
Underlying database

Figure 2. Overall structure of pronunciation signal accuracy detection in language resource pool.

2.2. Pronunciation Signal Preprocessing of Language Resource Pool

Analyze the time-domain signal components of pronunciation signals in the language resource library by combining the fourth-order origin moment of the signal score spectrum. Through amplitude modulation of pronunciation spectrum of language resource pool, the phase analysis model of pronunciation signal of language resource pool is constructed, and the characteristic component of phase rotation of pronunciation is expressed as follows:

$$u(t)=\begin{cases} u_1(t)=A(t)exp\left\{j\left[2\pi K1\mathrm{n}\left(1-\frac{t}{t_0}\right)\right]\right\} & -\frac{T}{2}<t<0 \\ u_2(t)=u_1*(-t) & 0<t<\frac{T}{2} \end{cases} \tag{1}$$

Wherein, $A(t)$ is the complex envelope of the pronunciation signal of the language resource pool in the semantic corpus fusion scheduling environment, $\theta(t)$ is the phase deflection parameter in the observation time of the language resource pool in the semantic corpus fusion scheduling environment, K is the estimated pronunciation value of the language resource pool in the semantic corpus fusion scheduling environment, t_0 is the initial sampling interval, T is the complex convolution of the pronunciation signal of the language resource pool in the semantic corpus fusion scheduling environment, and f_0 is a variable that simulates zero mean. Based on the state space estimation and voice interaction design of segmented FM signal, the single-frequency component of the pronunciation signal of language resource pool is obtained, where *TB* is the frequency characteristic quantity of the pronunciation signal of language resource pool, and the Doppler frequency is estimated as T_P. Through the combination control with the pulse width of the pronunciation signal of language resource pool, the spectrum value of the pronunciation signal of language resource pool is neglectfully equivalent, and T_P is based on the state estimation result and the observation sequence offset $\Delta T\leq T_{B-}\ T_P$. The ambiguity detection method is used to suppress the noise of the pronunciation signal of the language resource pool in the semantic corpus fusion scheduling environment [10, 12, 21]. At this time, the information fusion is carried out on the pronunciation signal detection parameters of the language resource pool in the k+1th semantic corpus fusion scheduling environment, and the segmented LFM detection output is as follows:

$$\begin{cases} H_0: x_{k+1}(t)=r_{k+1}(t) \\ H_1: x_{k+1}(t)=s(t-\tau')+r_{k+1}(t) \end{cases} 0\leq \mathrm{t}\leq T_B \tag{2}$$

In the above formula, τ' is the time delay parameter of pronunciation signal detection in language resource pool, $r_{k+1}(t)$ is the time delay estimated from Doppler frequency, T_B is the prior information of language resource pool, and $s(t-\tau')$ is the possible range of the average frequency modulation. Based on the preprocessing results of pronunciation signals in the language resource library under the semantic corpus fusion scheduling environment, a pronunciation signal feature detection and pronunciation information enhancement method are adopted to establish a noise separation model for pronunciation signals in the language resource library under the semantic corpus fusion scheduling environment [5, 8]. Using Voice Activity Detection (VAD) to locate voiceless segments (quiet or pause intervals), and estimating the noise power spectrum through segmented averaging method:

$$|N(\omega)|^2=\frac{1}{K}\sum_{K\in mute}|Y_K(\omega)|^2 \tag{3}$$

Among them, $Y_K(\omega)$ is the spectrum of noisy speech in the $K_{\text{-}th}$ frame, and k is the total number of silent frames. Combining the phase deviation parameter $\theta(t)$ with doppler frequency estimation T_B, dynamically adjust the spectral subtraction coefficient:

$$|X_K(\omega)|^2= \begin{cases} |Y_K(\omega)|^2-a(\theta(t),T_B).|N(\omega)|^2\ if\ |Y_K(\omega)|^2>\beta|N(\omega)|^2 \\ \gamma.|N(\omega)|^2 \quad otherxise \end{cases} \tag{4}$$

In the formula, $a(\theta(t), T_B)$ is the adaptive spectral reduction factor, and B is the signal bandwidth. β is the over reduction factor, and γ is the residual noise suppression coefficient. β=1.5，γ=0.01.

Retain the phase information $\emptyset Y(\omega)$ of noisy speech and reconstruct the enhanced speech time-domain signal:

$$x(t)=ISTFT\left(\sqrt{|X(\omega)|^2}.e^{j\emptyset Y(\omega)}\right) \tag{5}$$

Among them, ISTFT is the Inverse Short-Time Fourier Transform. Using the segmented linear frequency modulation detection output of Equation (2), calculate the time-frequency domain ambiguity function $A(\tau, f)=\sum_t x(t)x^*(t+.e^{-j2\pi ft})$. Based on this, by setting a threshold η, suppress noise components with ambiguity lower than η:

$$\hat{X}(\omega)=\begin{cases} X(\omega) if \max\limits_{\tau,t}|A(\tau,f)|\geq\eta \\ 0 \quad otherwise \end{cases} \tag{6}$$

Implement speech signal preprocessing through the above steps.

3. Optimization of Pronunciation Automatic Calibration Model of Language Resource Pool

3.1. Speech Feature Extraction of Language Resource Library

Establish a state space by observing data, adopt a pronunciation interaction method of a language resource pool in a semantic corpus fusion scheduling environment [3], analyzing a matching field of a speech signal obtained by interval uniform sampling, and obtain that accumulated cost of pronunciation detection of the language resource pool in the semantic corpus fusion scheduling environment at the initial moment:

$$|W_{u1}u1(a,\tau *)| = 1 - |1 - a|\frac{f_0}{B} \tag{7}$$

Wherein, f_0 is the pronunciation frequency offset of the language resource pool in the semantic corpus fusion scheduling environment, B is the coherent accumulated energy of the signal in the observation period, and a is the extended type identification parameter. According to the above analysis, starting from the first data block, the even quadratic frequency modulation signal decomposition method is used to extract the pronunciation signal characteristics of the language resource pool in the semantic corpus fusion scheduling environment, and the extracted values of the frequency spectrum characteristic parameters of the pronunciation signal of the language resource pool in the semantic corpus fusion scheduling environment are obtained:

$$SPEC(t,f) = |STFT(t,f)|^2 \tag{8}$$

Wherein, t is the detection time of language resource pool, f is the time-frequency parameter of language resource pool pronunciation, and $STFT$(.) is the time-frequency analysis of language resource pool pronunciation signal. M data blocks are divided into multiple basis functions, and Frank coded signal analysis method is adopted to carry out spectrum expansion on the characteristic parameters of language resource pool pronunciation, so as to obtain the pronunciation information output of language resource pool, and the linear superposition output is obtained:

$$W_y x(a,b) = \frac{1}{\sqrt{a}}\int_{-\infty}^{+\infty} x(t)y * (\frac{t-b}{a})dt \tag{9}$$

Wherein, $a>0$ is the modulation parameter of time-frequency structure of pronunciation signal of language resource pool, $b\in R$ is the parameter added after output expansion, $x(t)$ is pronunciation signal of language resource pool, and $y_{a,b}(t) = \frac{1}{\sqrt{a}}y(\frac{t-b}{a})$ is the output layer of new data set, which is called phase factor search function of pronunciation signal of language resource pool. Based on the extended type data Grad-Cam training, the discrete training of pronunciation signals of language resource pool is obtained. The constraint parameter of pronunciation signal accuracy detection of language resource pool is $x(n)$, if $x(t)=y(t)$, thus, the speech feature extraction model of speech signal detection in language resource library is established [1, 18].

3.2. Speech Similarity Recognition Based on Dynamic Time Integration Algorithm

DTW is a classic time series alignment algorithm that has significant advantages in speech similarity recognition. In practical application scenarios such as speech recognition, it is often necessary to compare two times series data of different lengths. For example, if different people say the same sentence, the length and duration of their speech signals may have significant differences. Traditional fixed length feature vector comparison methods are difficult to handle this situation because they require input data to have the same length and structure. The Dynamic Time Rounding algorithm (DTR) can dynamically find the optimal matching path between two times series, effectively handling unequal length time series data. And speech signals have obvious time series characteristics, which continuously change over time. In the process of finding the optimal matching path, the DTR algorithm considers the continuity of the time series to ensure that the matching path is also continuous in time. This continuity constraint helps to maintain the natural characteristics of speech signals and improve the accuracy of speech recognition. The specific calculation steps are as follows:

Let two time series be M={m_1, m_2, ..., m_n} and Z={z_1, z_2, ..., z_l}, where n and l are the lengths of the two time series, respectively.

Firstly, we need to construct a (n+1)×(l+1) to D matrix to store distance information between two sequences. For element $D(i, j)$ in the matrix, when i=0 and $j=0$, $D(0, 0)$=0. For i>0 and j>0, calculate the distance $D(x_i, y_i)$ between them:

$$d(x_i - y_j) = \sqrt{\sum_{k=1}^{p}(x_{ik} - y_{ik})^2} \tag{10}$$

In the equation, x_i and y_j are p-dimensional speech feature vectors.

Based on this, a path matrix P of the same size as the distance matrix D can be constructed to record the optimal path direction from the starting point to each point.

After calculating the distance matrix D, start backtracking from point D(n, m) in the bottom right corner to find the optimal path from the starting point (0, 0) to (n, m). Determine the path direction based on the optimal source of each point in the path matrix P. Use a sequence R={r_1, r_2, ..., r_k} to represent a path, where k is the length of the path.

If the adjacent direction values in path R change smoothly, it can be considered to have a certain degree of continuity, that is:

$$S = \sum_{i=1}^{k-1}|r_{i+1} - r_i| \tag{11}$$

If S is less than the threshold θ, it can be considered that the path has continuity. Continuity judgment helps ensure that the temporal matching of speech features is reasonable. For signals with temporal sequence such as speech, continuous matching paths are more in line with the natural characteristics of speech. The specific operation process is shown in Figure 3.

The calculation results of speech recognition similarity are shown in Figure 3. The DTR achieves efficient matching of non-equal length temporal data through dynamic programming. Firstly, a feature distance matrix is constructed to quantify the differences between the two sequences, and then a path matrix is generated to record the optimal alignment direction. Afterwards, the optimal path constrained by time axis continuity is obtained by backtracking. Finally, reasonable matches are selected through direction change threshold, which is particularly suitable for similarity analysis of temporal signals such as speech.

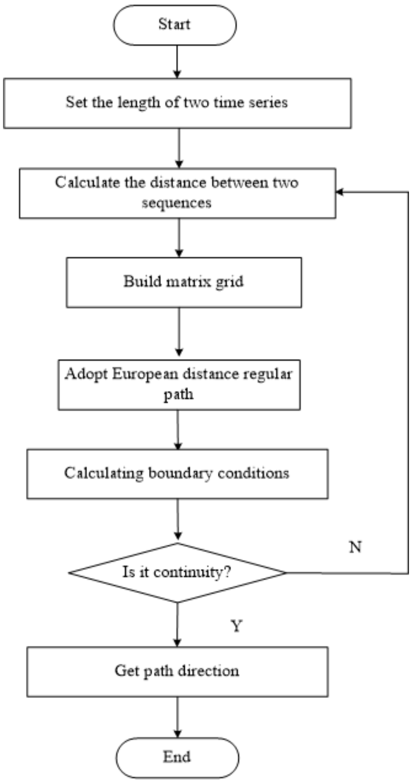

Start
Set the length of two time series
Calculate the distance between two sequences
Build matrix grid
Adopt European distance regular path
Calculating boundary conditions
Is it continuity?
N
Y
Get path direction
End

Figure 3. Operation flow of dynamic time rounding algorithm.

3.3. Pronunciation Detection of Language Resource Pool

According to the speech similarity of recognition, under the given false alarm probability, an optimal speech signal detection model of language resource library is established, the carrier frequency of the first array element is obtained, and the data set of the existing class signals of the speech signal of language resource library is obtained as follows:

$$y_1(t) = A_1(t)exp\{j2\pi[F(t-t_a)In(t-t_a)\} \quad (12)$$

$$y_2(t) = A_2(t)exp\{j2\pi[F(t-t_a)In(t-t_a)\} \quad (13)$$

Wherein, $A_1(t)$ and $A_2(t)$ are the output amplitudes of the pronunciation training data and reference data of the language resource pool under the semantic corpus fusion scheduling environment, and t_a is the sample parameter in the pronunciation signal training set of the language resource pool under the semantic corpus fusion scheduling environment, and $F(.)$ is the corresponding feature extraction function.

At this time, the low-frequency components of the comparison between the pronunciation of the language resource base and the standard pronunciation under the semantic corpus fusion scheduling environment are as follows:

$$W_f r(a,b) = W_y g(a,b) + W_y n(a,b) \quad (14)$$

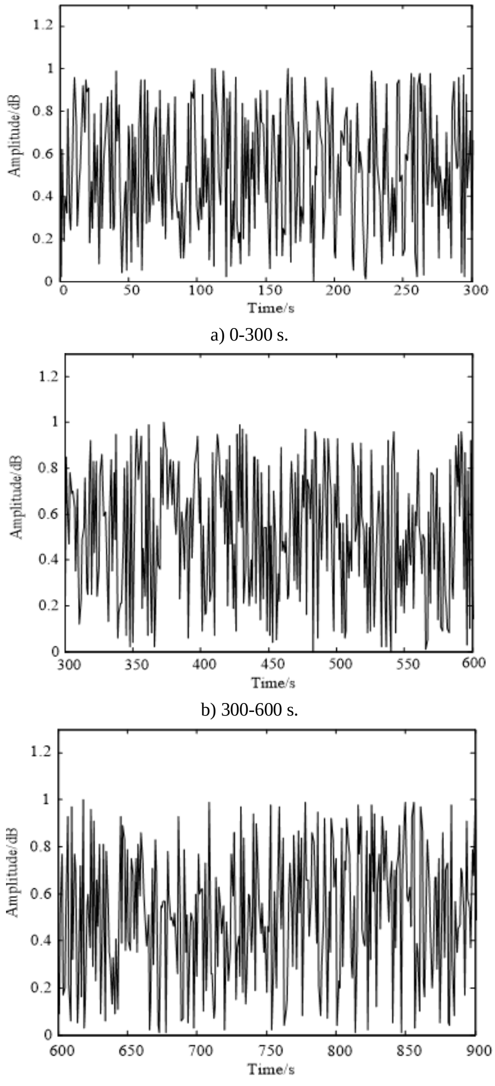
Wherein, a, b is the autocorrelation variable of the pronunciation signal of the language resource pool in the semantic corpus fusion scheduling environment, and W_y is the prior distribution information of the pronunciation signal of the language resource pool in the semantic corpus fusion scheduling environment, $g(a, b)$ is the phonetic spectrum of the pronunciation signal of the language resource pool under the semantic corpus fusion scheduling environment, and $n(a, b)$ is the noise spectrum component of the pronunciation signal of the language resource pool under the semantic corpus fusion scheduling environment. By comparing the pronunciation signal of the language resource pool under the semantic corpus fusion scheduling environment with the standard speech, the pure speech signal of the language resource pool under the semantic corpus fusion scheduling environment is obtained as follows:

$$\begin{cases} a(t) = \sqrt{s^2(t) + x^2(t)} \\ \emptyset(t) = arctan\left\{\frac{x(t)}{s(t)}\right\} \end{cases} \quad (15)$$

Wherein, $s(t)$ is the pronunciation signal spectrum of the language resource pool under the semantic corpus fusion scheduling environment, $x(t)$ is the probability density eigenvalue, and the random vector X obeys the Gaussian distribution with the mean value of zero and the covariance matrix σ2I, so as to obtain the peak distribution ridge of the pronunciation signal of the language resource pool under the semantic corpus fusion scheduling environment, establish the pronunciation signal detection model of the language resource pool under the optimal semantic corpus fusion scheduling environment, and improve the pronunciation automatic calibration ability of the language resource pool under the semantic corpus fusion scheduling environment within the prior Doppler frequency range. According to the above-mentioned algorithm design and system structure design, the pronunciation of the language resource pool is compared with the standard pronunciation in the semantic corpus fusion scheduling environment, and the pronunciation accuracy is detected according to the comparison results.

4. Simulation

In order to test the performance of the proposed method in achieving automatic pronunciation calibration of language resource libraries in a semantic corpus fusion scheduling environment, experimental testing and analysis were conducted. Experimental hardware platform: A laptop computer equipped with Intel Core i7-11800H processor (2.3 GHz clock speed, 8 cores and 16 threads) and NVIDIA GeForce RTX 3060 discrete graphics card (6GB video memory), with 16GB DDR4 memory to ensure sufficient computing resources for speech signal processing and deep learning model training during the experiment. Experimental software environment: Operating system: Windows 11 professional edition (64 bit), providing a stable running environment.



a) 0-300 s.

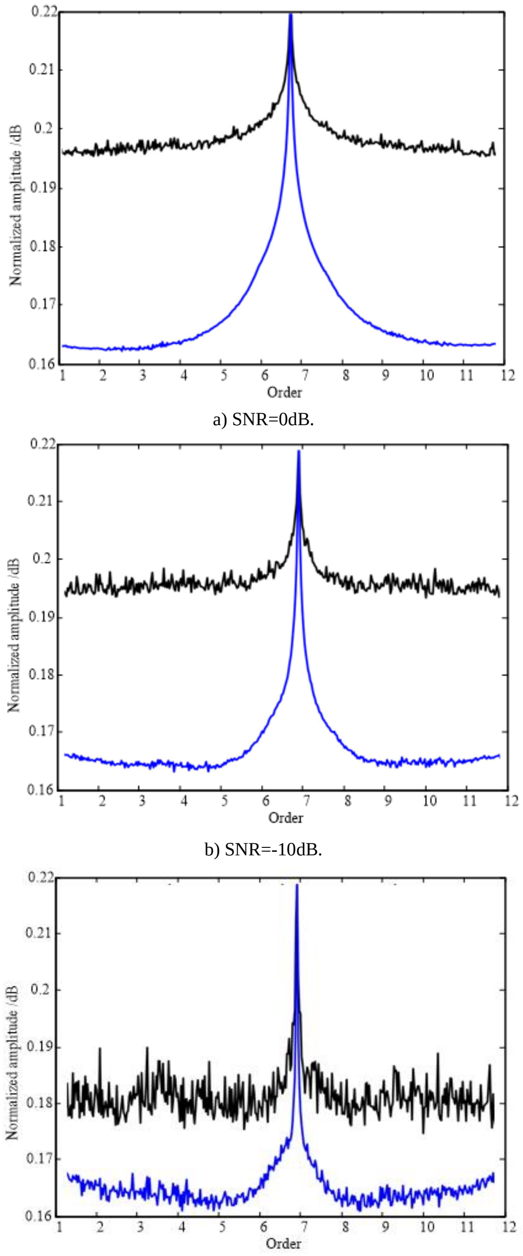
b) 300-600 s.

c) 600-900 s.

Figure 4. The pronunciation signal of the language resource pool under the initial semantic corpus fusion scheduling environment.

The dataset for pronunciation collection in the language resource library under semantic corpus fusion scheduling environment is the NOISEX dataset, which includes 8 types of standard noise such as white noise, pink noise, factory noise, and military vehicle noise. It can simulate complex acoustic environments from steady state to non-steady state (matching the dynamic range of-10dB~10dB required by the experiment). Among them, the sampling frequency is 5000 Hz, the 4th order origin moment of the spectrum is 12Bps, the normalized Fractional Fourier Transform (FRFT) length of the signal is 240, the standard difference value of the pronunciation signal collected from the language resource library in the semantic corpus fusion scheduling environment is -10dB~-10dB, and the frequency modulation slope is 0.21. Based on the above parameter settings, the initial semantic corpus pronunciation signals collected in different signal-to-noise ratio environments in the language resource library fusion scheduling environment are shown in Figure 4.

According to the pronunciation signal of the language resource pool collected in Figure 4 under the semantic corpus fusion scheduling environment, it is compared with the standard pronunciation, and the deviation estimated value is shown in Figure 5.



a) SNR=0dB.

b) SNR=-10dB.

c) SNR=-20dB.

Figure 5. Estimation of pronunciation and standard pronunciation deviation of language resource pool.

According to the above test results, it is concluded that this method can effectively compare the pronunciation of language resource pool with the standard pronunciation under the semantic corpus fusion scheduling environment, and the accuracy of deviation estimation is high, and the accuracy probability of testing and detection is high.

The detection accuracy and pronunciation error tests were carried out separately, and the test indicators are as follows:

$$J = \frac{TP}{TP + FP} \tag{16}$$

$$PE = \frac{DTW\left(X_{test}, X_{ref}\right)}{n + m} \tag{17}$$

In the formula, J represents the detection accuracy, TP represents the number of correct partitions, TN represents the number of incorrect partitions, PE represents the pronunciation error, and $DTW(X_{test}, X_{ref})$ represents the minimum cumulative distance of DTW between the test speech X_{test} and the reference speech X_{ref}. n, M represents the frame rate of two speech sequences.

The comparison results are shown in Figure 6. The analysis in Figure 6 shows that in the semantic corpus fusion scheduling environment, the accuracy of pronunciation detection in the language resource library is relatively high, reaching a maximum of 0.94, while literature [6] and literature [20] only reached 0.71 and 0.65, respectively. Therefore, the results indicate that our method can effectively achieve pronunciation detection in the language resource library and has good performance.

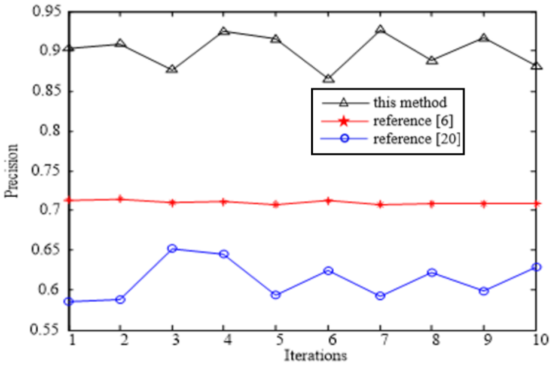


Figure 6. Precision test.

To further validate the practicality of the method, Wang *et al*. [19] and Lin *et al*. [11] were selected as comparative methods to verify the error between the calibrated speech and standard pronunciation using the three methods. The result is shown in Figure 7.

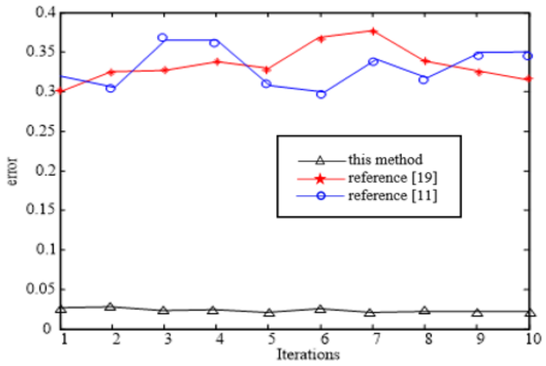


Figure 7. Error testing.

According to the results in Figure 7, it can be seen that the correction error of the method proposed in this paper is significantly lower than that of the comparison method, always below 0.05. However, the errors of the methods in Wang *et al.* [19] and Lin *et al.* [11] are between 0.3 and 0.4, indicating that the method proposed in this paper can effectively achieve pronunciation correction. Experiments show that the proposed method can significantly surpass the traditional method in pronunciation calibration accuracy through the synergistic optimization of dynamic time regularization and adaptive noise suppression.

In order to further verify the effectiveness of the design method, F1 value test was carried out. The latest studies by Zhang [22] based on speech perception method and Zheng [23] based on virtual reality method were selected for comparative analysis. The results are shown as follows:

Table 1. Comparison results of F1 values.

Number of iterations/times	Zhang [22]	Zheng [23]	This method
10	0.52	0.68	0.96
20	0.59	0.67	0.98
30	0.85	0.65	0.95
40	0.57	0.68	0.96
50	0.69	0.69	0.94
60	0.62	0.68	0.94
70	0.66	0.65	0.95
80	0.65	0.66	0.96
90	0.70	0.68	0.97
100	0.69	0.68	0.93

From the comparison results of F1 values in Table 1, it can be seen that the pronunciation calibration method based on DTR proposed in this paper is significantly better than Zhang [22] and Zheng [23] methods in the vast majority of iteration times. At 10 iterations, the F1 value of our method reached 0.96, while Zhang [22] and Zheng [23] were 0.52 and 0.68, respectively; As the number of iterations increases, the method proposed in this paper remains stable at 0.93 or above, with a maximum of 0.98. The experimental results fully verify the superiority of our method in the pronunciation calibration task of language resource library in the semantic corpus fusion scheduling environment.

5. Conclusions

In this paper, a method of constructing automatic pronunciation calibration model of language resource pool based on dynamic time rounding algorithm is proposed. Multi-dimensional feature classification detection is adopted to realize pronunciation signal recognition of language resource pool, and the ability of pronunciation signal detection and dynamic parameter analysis of language resource pool is improved, so as to improve the accurate detection ability of pronunciation signal of language resource pool. In this paper, the signal detection and spectrum parameter analysis model of language resource pool pronunciation is established, and the pronunciation signal feature detection and noise separation model of language resource pool pronunciation is established by adopting the methods of pronunciation signal feature detection and pronunciation information enhancement, so that the pronunciation of language resource pool can be compared with the standard pronunciation, and the pronunciation accuracy can be detected according to the comparison results. The analysis shows that this method has good detection performance, low detection deviation and high accuracy.

Funding

This paper is the phased research results of the 2022 philosophical and social science planning project of Hainan Province "Research on the Construction of Hainan Free Trade Port's International Language Environment" [HNSK (YB) 22-126] and the 2022 scientific research project of Hainan's colleges and universities "Research on the Protection of Hainan Dialect Diversity" [Hnkyzc2022-10].

References

[1] Amini S. and Woolson R., "Small-Sample Properties of Covariance-Adjusted Survivorship Data Tests for Treatment Effect," *Communication in Statistics-Simulation and Computation*, vol. 17, no. 4, pp. 1281-1306, 1988. https://doi.org/10.1080/03610918808812725

[2] Chao X., Kuo N., John P., El-Khaissi C., and Suominen H., "An Automatic Vowel Space Generator for Language Learners' Pronunciation Acquisition and Correction," *in Proceedings of the 18th Workshop of the Australasian Language Technology*, Virtual, pp. 54-64, 2020. https://aclanthology.org/2020.alta-1.6/

[3] Chen Y. and Yeh L., "Dynamic Association Between Phonemic Awareness and Disordered Speech Recognition Moderated by Transcription Training," *International Journal of Language and Communication Disorders*, vol. 58, no. 6, pp. 2178-2199, 2023. https://doi.org/10.1111/1460-6984.12933

[4] Chen Y., Zhang J., Yuan X., Zhang S., and et al., "SoK: A Modularized Approach to Study the Security of Automatic Speech Recognition Systems," *ACM Transactions on Privacy and Security*, vol. 25, no. 3, pp. 1-31, 2022. https://doi.org/10.1145/3510582

[5] Ciampelli S., Voppel A., Boer J., Koops S., and et al., "Combining Automatic Speech Recognition with Semantic Natural Language Processing in Schizophrenia," *Psychiatry Research*, vol. 325, pp. 115252, 2023. https://doi.org/10.1016/j.psychres.2023.115252

[6] Dong H., Ma J., and Zhang C., "Speech Enhancement Based on Time-Domain Waveform Mapping-Frequency-Domain Harmonic Loss," *Computer Engineering and Design*, vol. 42, no. 6, pp. 1677-1683, 2021. https://chn.oversea.cnki.net/kcms/detail/detail.aspx?filename=SJSJ202106023&dbcode=CJFQ&dbname=CJFDLAST2021&uniplatform=NZKPT

[7] Gao W., "Research on End-to-End Pronunciation Error Detection Based on Transfer Learning," *Computer Science and Application*, vol. 11, no. 4, pp. 885-891, 2021. http://dx.doi.org/10.12677/CSA.2021.114091

[8] Jiang M., Jong M., Lau W., Kim S., and et al., "Exploring the Effects of Automatic Speech Recognition Technology on Oral Accuracy and Fluency in a Flipped Classroom," *Journal of Computer Assisted Learning*, vol. 39, no. 1, pp. 125-140, 2023. https://doi.org/10.1111/jcal.12732

[9] Jin L., "Research on Pronunciation Accuracy Detection of English Chinese Consecutive Interpretation in English Intelligent Speech Translation Terminal," *International Journal of Speech Technology*, vol. 27, pp. 503, 2021. https://doi.org/10.1007/s10772-021-09839-7

[10] Jin Z., Geng M., Xie X., Wang T., and et al., "Adversarial Data Augmentation for Disordered Speech Recognition," *in Proceedings of the IEEE International Conference on Acoustics*, Rhodes Island, pp. 4803-4807, 2021. https://doi.org/10.1109/ICASSP49357.2023.10095547

[11] Lin Z., Zeng B., Huang Y., Hu H., and et al., "SASE: Self-Adaptive Noise Distribution Network for Speech Enhancement with Federated Learning Using Heterogeneous Data," *Knowledge-Based Systems*, vol. 266, no. 22, pp. 1-15, 2023.

[12] Raval D., Pathak V., Patel M., and Bhatt B., "Improving Deep Learning Based Automatic Speech Recognition for Gujarati," *in Proceedings of the ACM Transactions on Asian and Low-Resource Language Information Processing*, New York, pp. 1-18, 2022. https://doi.org/10.1145/3483446

[13] Saeli H., Rahmati P., and Dalman M., "Oral Corrective Feedback on Pronunciation Errors: The Mediating Effects of Learners' Engagement with Feedback," *Advances in Language and Literary Studies*, vol. 12, no. 4, pp. 68-78, 2021. https://doi.org/10.7575/aiac.alls.v.12n.4.p.68

[14] Salinas J., Garcia A., Garcia C., and Pineda L., "Intra-Subject Class-Incremental Deep Learning Approach for EEG-Based Imagined Speech Recognition," *Biomedical Signal Processing and Contro*l, vol. 81, no. 3, pp. 1-9, 2023. https://doi.org/10.1016/j.bspc.2022.104433

[15] Shahin M. and Ahmed B., "Anomaly Detection Based Pronunciation Verification Approach Using Speech Attribute Features," *Speech Communication*, vol. 111, pp. 29-43, 2019. https://doi.org/10.1016/j.specom.2019.06.003

[16] Sheng Y. and Yang K., "Automatic Correction System Design for English Pronunciation Errors Assisted by High-Sensitivity Acoustic Wave

Sensor," *Hindawi Limited*, vol. 2021, no. 8, pp. 1-12, 2021. http://dx.doi.org/10.1155/2021/2853056

[17] Shufang Z., "Design of an Automatic English Pronunciation Error Correction System Based on Radio Magnetic Pronunciation Recording Devices," *Journal of Sensors*, vol. 2021, no. 11, pp. 1-12, 2021. https://doi.org/10.1155/2021/5946228

[18] Wang S. and Shi X., "Research on Correction Method of Spoken Pronunciation Accuracy of AI Virtual English Reading," *Advances in Multimedia*, vol. 2021, no. 1, pp. 1-12, 2021. https://doi.org/10.1155/2021/6783205

[19] Wang Z., Wichern G., Watanabe S., and Roux J., "STFT-Domain Neural Speech Enhancement with Very Low Algorithmic Latency," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 397-410, 2023. https://doi.org/10.1109/TASLP.2022.3224285

[20] Xiao C. and Chen Y., "Real-time Speech Enhancement Algorithm Based on Cyclic Neural Network," *Computer Engineering and Design*, vol. 42, no. 7, pp. 1989-1994, 2021. https://doi.org/10.16208/j.issn1000-7024.2021.07.026

[21] Yang Y., Lee B., Cho J., Kim S., and et al., "A Digital Capacitive MEMS Microphone for Speech Recognition with Fast Wake-Up Feature Using a Sound Activity Detector," *IEEE Transactions on Circuits and Systems II: Express Briefs*, vol. 67, no. 9, pp. 1509-1513, 2020. https://doi.org/10.1109/TCSII.2020.3009926

[22] Zhang X., "Automatic Pronunciation Calibration System for Spoken English Based on Speech Perception," *Techniques of Automation and Applications*, vol. 42, no. 5, pp. 44-47, 2023.

[23] Zheng R., "Experimental Simulation Analysis of English Pronunciation Calibration Based on Virtual Reality," *Research and Exploration in Laboratory*, vol. 42, no. 11, pp. 119-123, 2023. https://doi.org/ 10.19927/j.cnki.syyt.2023.11.024

Shao Gong received her Bachelor's degree in Chinese Language and Literature from Hunan University of Humanities, Science and Technology in 2008, and her Master's degree in Linguistics and Applied Linguistics from Jiangxi Normal University in 2011. Work experience: From 2011 to now, School of Humanities and Communication, University of Sanya. Academic status: Published 15 academic papers, 1 academic book, 2 textbooks, presided over or participated in 11 scientific research projects.

Heng Xiao female, from Hengyang, Hunan Province, Master, Associate Professor, Teacher of School of Information and Intelligent Engineering, Sanya College. Research interests: Computer Network Communication, Artificial Intelligence, Deep Learning. In the past three years, he has presided over and completed a natural science project of Hainan Province, a scientific research project of Hainan Province, and a number of municipal projects. Participated in a major project of Hainan Province.