Hindawi
Security and Communication Networks
Volume 2023, Article ID 9760525, 1 page
https://doi.org/10.1155/2023/9760525

Retraction

Retracted: Modeling and Simulation of English Speech Optimization Teaching Recognition Based on Intelligent Edge Detection Algorithm

Security and Communication Networks

Received 20 June 2023; Accepted 20 June 2023; Published 21 June 2023

This article has been retracted by Hindawi following an investigation undertaken by the publisher [1]. This investigation has uncovered evidence of one or more of the following indicators of systematic manipulation of the publication process:

(1) Discrepancies in scope
(2) Discrepancies in the description of the research reported
(3) Discrepancies between the availability of data and the research described
(4) Inappropriate citations
(5) Incoherent, meaningless and/or irrelevant content included in the article
(6) Peer-review manipulation

The presence of these indicators undermines our confidence in the integrity of the article's content and we cannot, therefore, vouch for its reliability. Please note that this notice is intended solely to alert readers that the content of this article is unreliable. We have not investigated whether authors were aware of or involved in the systematic manipulation of the publication process.

Wiley and Hindawi regrets that the usual quality checks did not identify these issues before publication and have since put additional measures in place to safeguard research integrity.

We wish to credit our own Research Integrity and Research Publishing teams and anonymous and named external researchers and research integrity experts for contributing to this investigation.

The corresponding author, as the representative of all authors, has been given the opportunity to register their agreement or disagreement to this retraction. We have kept a record of any response received.

References

[1] Y. Tang and X. Tian, "Modeling and Simulation of English Speech Optimization Teaching Recognition Based on Intelligent Edge Detection Algorithm," *Security and Communication Networks*, vol. 2022, Article ID 9314068, 12 pages, 2022.

Hindawi
Security and Communication Networks
Volume 2022, Article ID 9314068, 12 pages
https://doi.org/10.1155/2022/9314068

Research Article

Modeling and Simulation of English Speech Optimization Teaching Recognition Based on Intelligent Edge Detection Algorithm

Yuan Tang[1] **and Xin Tian**[2]

[1]*School of Humanities and Arts, Jiaxing Nanhu University, Jiaxing 314000, Zhejiang, China*
[2]*School of Finance, Dalian University of Finance and Economy, Dalian 116600, Liaoning, China*

Correspondence should be addressed to Xin Tian; tianxin0514@dlufe.edu.cn

Received 13 April 2022; Revised 31 May 2022; Accepted 16 June 2022; Published 28 June 2022

Academic Editor: Mohammad Ayoub Khan

English speech modeling is one of the key problems in the field of speech recognition. Its accuracy directly affects the performance of the English speech recognition system, and how to establish a more accurate acoustic model has always been the focus of researchers. This paper is based on the intelligent edge detection algorithm of the English speech optimization teaching recognition modeling simulation analysis to improve the accuracy of acoustic models such as parameters and the performance of continuous speech recognition system as the main purpose. In this paper, the accuracy of the neural network is improved on the premise of improving the training speed and decoding speed of the model. It proposes a new model and how to use the intelligent edge detection algorithm to build a complete English speech optimization teaching recognition system. The whole system includes the mobile terminal and the server, which realizes the most basic business logic of the speech recognition system. The experimental results of this paper show that from the point of view of the average recognition rate, the recognition effect of the optimized feature set has been further improved compared with the fusion feature. From the point of view of the recognition rate under different SNR environments, the recognition rate of the optimized feature set PCA-Features2 decreased by 0.47% compared with the FFPRLS_D + TEOCC feature set under 10dB10 words. Compared with the FFPLMS_D + TEOCC feature set, the recognition rate under 5dB10 words also drops by 0.47%.

1. Introduction

English is the most widely spoken international language in the world and is increasingly valued as a second language by learners (English as Second Language (ESL)). However, due to cultural, regional, and lifestyle differences, ESL learners face a variety of challenges when learning English, including listening, speaking, reading, and writing, of which the most important and difficult is the teaching of English phonetics, and phonetic grammar errors are the most common error types in English phonetics teaching. Speech recognition is equivalent to the auditory system of the machine, taking the speech signal as the research object, and allowing the computer to understand the human language.

Due to the nonstationary time-varying characteristics of speech signals, the Fourier transform is used to analyze the frequency domain information of speech in the traditional feature extraction process. Practice shows that the features that conform to the auditory characteristics of the human ear have a positive effect on the improvement of speech recognition. It can enhance the noise robustness of the recognition system. Therefore, the study of auditory features has important theoretical significance in the field of anti-noise speech recognition.

The innovation of this paper is as follows: (1) The energy feature TEOCC is added, and the constructed fusion feature set combines the human hearing characteristics and energy characteristics and has better speech recognition performance. (2) Based on the problem of feature redundancy in

the fusion feature set, a feature optimization method based on PCA is proposed. (3) The design experiment compares and analyzes the three optimized fusion feature sets to obtain the optimal feature set, so as to better characterize the complete characteristics of the speech signal and improve the overall recognition rate of the recognition network.

2. Related Work

In audio streams containing multiple speakers, speaker classification helps determine "who is speaking when." In this work, Subba Ramaiah and Rajeswara Rao proposed a novel speaker classification system. The system uses Tangent Weighted Mel Frequency Cepstral Coefficients (TMFCC) as characteristic parameters and Lion algorithm. It is used to cluster audio streams detected by voice activity into specific speaker groups [1]. Misirov stipulated the teaching of English pronunciation in primary and secondary schools (grades), the teaching of English approximate pronunciation, the selection of English pronunciation teaching materials in primary schools, and the teaching methods and skills of English pronunciation in primary schools [2]. The goal of pronunciation teaching has shifted from unaccented or native-like pronunciation to intelligibility. Teaching practices for nonnative English speakers vary and are often based on teacher input rather than research results. Vančová's research found that a good theoretical background of teachers can improve students' awareness and overall performance of pronunciation phenomena, whether at the level of discourse or supradiscourse [3]. Gashaw examined the pitch and rhythmicity of Ethiopian English speakers' pronunciation to determine the presence of syllabic or tonal rhythm in English speech. The results showed that native samples of Amharic showed true peaks in almost all words that required more pronunciation [4]. Difficulties in producing weak/loose vowels were highlighted during the exam. Chika et al. analysis of the correlation between overall goodness, and each of the other indicators shows that improving both segmental and prosodic features is crucial for Japanese learners to achieve good pronunciation. Using multiple indicators for systematic pronunciation assessment is not only conducive to a deeper and broader understanding of Japanese and English but also to the development of pronunciation teaching [5]. The results of Kolesnikova's study of American English pronunciation could be used as a starting point for diagnosis, post hoc validation, and adjustment of possible misreadings [6]. The purpose of Han et al. research was to propose a visual teaching method for the pronunciation of English vowels, especially for hearing-impaired people who mainly rely on visual aids, based on support vector machine technology. The lip shape of each vowel is improved by using SVM technology to extract speech features from sounds that are hard to hear by the ear. The advantage of vowel labial refinement is that language learners can easily see the movements of the vocal organs with their eyes. This is beneficial for the hearing impaired to learn and teach English vowels [7]. Shadowing has been practiced in Japanese English classes for decades. Numerous studies have confirmed its effectiveness in improving learners' listening and pronunciation skills. Therefore, Zajdler will base on the auditory and cognitive basis of shadows and then propose a classification of shadow types. Finally, it examines the practical aspects of shadowing as an effective classroom CFL instructional technique [8]. Although they are effective in improving learners' listening and pronunciation skills, there is still a certain gap in the experimental results.

3. Speech Recognition Method of Edge Detection Algorithm

3.1. Problems Existing in the Recognition of English Pronunciation Optimization Teaching. In recent years, speech recognition technology has achieved great breakthroughs and practical results. However, there are still many scientific research problems in its technical field that need to be solved urgently [9–11]. As a research hot spot in the field of artificial intelligence, speech recognition technology should not be limited to a large number of theoretical studies but should also be applied to people's lives. The current difficulties of speech recognition technology are reflected in:

3.1.1. Establishing a Unified Voice Database. There are many kinds of speech libraries currently used, the common ones are as follows: TIDigits nonspecific digital string speech database, UCI ISOLET alphabet speech database, and UCI Vowel English vowel speech database. However, the data samples of these speech databases are limited. Because there are many types of speech, it is difficult to get close to life in terms of semantics and speech selection [12, 13]. This leads to the lack of generality and wide applicability in the research of speech databases. Therefore, the establishment of a unified standard, systematic, and complete high-quality speech database is the primary problem to be solved by speech recognition technology so far.

3.1.2. The Voice Signal Is Unstable and Contains a Large Amount of Information. The speech feature parameters are extracted from the speech of different speakers. Speech is unstable and time-varying and usually changes with the speaker's vocalization state, emotional changes, and physical conditions [14–16]. Due to the diversity of human pronunciation, the semantic information contained in different pronunciation states of the same person is different. Therefore, how to extract stable feature parameters that can fully characterize most of the useful information in speech signals is a key issue in the field of speech recognition research.

3.1.3. Establish a Feature Set That Can Fully Represent Most of the Effective Information in Speech. As the input of the speech classification model, the front-end features should contain most of the effective information of the speech signal, weaken the speaker's personality information, and have robust and stable performance [17, 18]. A single

feature is not enough to characterize the complete characteristics of the speech signal, so researchers describe the characteristics of the speech signal from different perspectives and fuse different speech features. Although the fusion of different types of features can improve the performance of the recognition network, it is also prone to feature redundancy, which weakens the ability of some features to represent information, so that the speech characteristics cannot be fully described [19]. Therefore, how to establish a feature set that can fully represent speech information is also one of the difficult problems faced by speech recognition.

3.2. RBM Model. A restricted Boltzmann machine (RBM) is a Markov random field with a special structure that can be viewed as an unsupervised learning process. It models the probability distribution of input patterns fixed at the explicit layer through this learning process [20]. V corresponds to the dimension of the observation vector, and the size of H is variable, which determines the complexity of the RBM model.

$$\begin{aligned} E(v,h,\theta) &= -v^T W h - a^T v - b^T h \\ &= -\sum_{i=1}^{V} v_i a_i - \sum_{j=1}^{H} h_i b_j - \sum_{i=1}^{V}\sum_{j=1}^{H} W_{ij} V_i h_j, \end{aligned} \tag{1}$$

Here, the RBM model parameters are $\theta = \{W,a,b\}$. W_{ij} represents the weight between the i-th node in the explicit layer and the jth node in the hidden layer. a_i and b_i represent the bias size of node i in the visible layer and node j in the hidden layer, respectively [21]. According to the Gibbs distribution, the probability that the RBM chooses to be in the current state (v, h) can be obtained as:

$$\begin{aligned} P(v,h;\theta) &= \frac{1}{Z}\exp(-E(v,h;\theta)) \\ Z &= \sum_{v}\sum_{h}\exp(-E(v,h;\theta)). \end{aligned} \tag{2}$$

This probability can be considered as the joint probability distribution of the explicit state and the hidden state. It is obtained by exponentially normalizing the energy of the RBM in the current state through the energy of the RBM in all possible states [22]. Z is the partition function, a regular term that takes into account the energy of the RBM in all states. Therefore, the marginal distribution of the explicit state vector can be derived from the joint distribution above:

$$P(v;\theta) = \frac{1}{Z}\sum_{h}\exp(-E(v,h;\theta)). \tag{3}$$

So far, the energy function can be obtained by combining the model parameters of the RBM with the state of its explicit and hidden layers, and finally, the probability distribution of the explicit state is derived [23]. Therefore, RBM is an energy-based probability distribution model.

The density model derived from the above energy function corresponds to the case where the explicit layer data follow a Bernoulli distribution. In speech recognition, the observation data of speech signal are a continuous value. Therefore, the explicit layer is changed to Gaussian distribution, and the Bernoulli distribution of the hidden layer remained unchanged, for this Gauss-Bernoulli type of RBM, and its energy function in state (y, h) is calculated as follows:

$$E(v,h;\theta) = \sum_{i=1}^{V}\frac{(v_i - a_i)^2}{2\sigma_i^2} - \sum_{j=1}^{H} h_j b_j - \sum_{i=1}^{V}\sum_{j=1}^{H} W_{ij} h_j \frac{v_i}{\delta_i}. \tag{4}$$

In practice, for the convenience of calculation, the variance of each node in the display layer is limited to 1. The training data are therefore normalized to a standard Gaussian distribution with zero mean and unit variance. In this way, the energy function of the RBRBM corresponding to the state (U, h) is simplified to:

$$E(v,h;\theta) = (x-a)^T(v-a) - b^T h - v^T W h. \tag{5}$$

Similar to Bernoulli RBM, the joint distribution of v and h can be derived, and then the marginal probability density distribution describing the GBRBM explicit layer can be obtained as:

$$\begin{aligned} P(v;\theta) &= \frac{1}{Z}\sum_{h}(-E(v,h;\theta)) \\ &= \frac{1}{Z}\sum_{h}\exp\left(-(v-a)^T(v-a) + b^T h + v^T W h\right) \\ &= \frac{1}{Z}\exp\left(-(v-a)^T(v-a)\right)\prod_{j=1}^{H}\sum_{h_j\in\{0,1\}}\exp\left(b_j h_j + v^T W h_j\right) \\ &= \frac{1}{Z}\exp\left(-(v-a)^T(v-a)\right)\prod_{j=1}^{H}\left(1+\exp\left(b_j + v^T w_j\right)\right). \end{aligned} \tag{6}$$

Among them, w_j is the jth column of the weight matrix w, and Z is the partition function:

$$Z = \int_v \sum_h \exp(-E(v,h;\theta))\mathrm{d}v. \tag{7}$$

3.3. DNN-HMM Model. Using the RBM pretraining method and BP algorithm in the previous two subsections, after the final DNN model is obtained through training, it can be used to accurately model the speech feature parameters. When a $(L+1)$ layer DNN is used to compute the posterior probability of the HMM bound state s given the acoustic observation vector o, the posterior probability is $P_{s|o}(s|o)$. The previous L layer, $l = \ldots L-1$, uses the sigmoid non-linear activation function to calculate the output of the current layer, which is also the input of the next layer. The topmost L layer uses the sofmax operation to calculate the posterior probability of all state classes:

$$p_{hj|v}^{\zeta}\left(h_j^{\zeta} \mid v^{\zeta}\right) = \frac{1}{\left(1 + e^{-z_j^{\zeta}\left(v^{\zeta}\right)}\right)} = \sigma\left(z_j^{\zeta}\left(v^{\zeta}\right)\right)$$

$$p_{s|v}^{\zeta}\left(s \mid v^{L}\right) = \frac{e^{z_s^{L}\left(v^{L}\right)}}{\left(\sum_{s} e^{-z_s^{L}\left(v^{L}\right)}\right)} = \mathrm{softmax}_s\left(z^{L}\left(v^{L}\right)\right) \tag{8}$$

$$z^{\zeta}\left(v^{\zeta}\right) = \left(W^{\zeta}\right)^{T} v^{\zeta} + a^{\zeta}.$$

Here, W^{ζ} and a^{ζ} represent the weight matrix and bias vector of the hidden layer ζ. h_j^{ζ} and $-z_j^{\zeta}(v^{\zeta})$ are the jth elements of h^{ζ} and $z^{\zeta}(v^{\zeta})$, respectively. DNN acoustic modeling uses stochastic gradient descent to update the weights in the process of error back-propagation:

$$\left(W^{\zeta}, a^{\zeta}\right) \leftarrow \left(W^{\zeta}, a^{\zeta}\right) + \eta \frac{\partial L}{\partial\left(W^{\zeta}, a^{\zeta}\right)}. \tag{9}$$

Here L represents an objective function to be optimized, and η represents the learning rate. In automatic speech recognition, the objective function is usually set as the total log posterior probability of all training modes $o(t)$ under their corresponding class labels $s(t)$, namely:

$$L = \sum_{t=1}^{T} \log\, P_{s|o}(s(t) \mid o(t)). \tag{10}$$

It is converted into a regularized likelihood, which can then be used by the Viterbi decoder to find the optimal path. This is the embodiment of the neural network Hybrid method.

$$\frac{p_m\left(o_t \mid q_t = i, s_t = j\right)}{p_m\left(o_t\right)} = \frac{p_m\left(o_t \mid q_t = i, s_t = j \mid o_t\right)}{p\left(q_t = i, s_t = j\right)}. \tag{11}$$

The above formula is the essence of DNN-HMM acoustic modeling. It uses the output layer of the DNN to directly model each binding state of the HMM, estimates the posterior probability distribution of each state, and obtains the corresponding regular likelihood value for decoding.

4. Design of the Recognition System for English Speech Optimization Teaching

In this chapter, an English speech optimization teaching recognition system will be built using the acoustic model proposed above. This system uses HMM with the acoustic model proposed above and the n-gram language model mentioned above and decodes it through a pregenerated static finite-state transition machine HCLG.fst.

4.1. System Architecture Design. This section will introduce the architectural design of the online language recognition system in detail. The system architecture is divided into two parts: physical architecture and logical architecture. A reasonable physical architecture can ensure the robustness and stability of the system. A good system logic architecture requires covering all the services of the system, ensuring high cohesion within modules and low coupling between modules.

4.1.1. System Physical Architecture. The speech recognition system has the characteristics of real-time and high concurrency. Therefore, a single-node server cannot meet the basic needs of the system [24–26]. The online speech recognition system uses a server cluster architecture and deploys the speech recognition server on each node in the cluster. It distributes tasks to computing nodes using load balancing. The node uses multithreaded concurrent decoding to minimize the feedback time. The topology of the physical architecture used in this system is shown in Figure 1:

As shown in Figure 1, the advantages of using a clustered architecture are as follows:

(1) High reliability. A cluster contains multiple compute nodes. Through load balancing, the pressure of each computing node in the cluster can be dynamically adjusted to avoid the accumulation of too many tasks on one server. For the user side, the time required to get feedback can be effectively reduced. For the server, it can avoid stopping the service due to server downtime caused by too many tasks.

(2) Strong scalability and maintainability. Using a cluster architecture, computing nodes can dynamically join or leave the cluster. When the system needs to be maintained or upgraded, related operations can be performed on the computing nodes one by one. When one computing node is maintained, other nodes continue to provide services, so as to avoid the temporary shutdown of services due to system maintenance or upgrades.

4.1.2. System Logical Architecture. The online speech recognition system adopts the architecture of a mobile terminal and server, which is divided into a presentation layer and a business logic layer. Different from the traditional three-tier architecture, this system has no online voice collection function. It only provides speech recognition services for users, so there is no data access layer in the logical architecture of this system, and it does not involve data access. The logical architecture of the system is shown in Figure 2.

As shown in Figure 2, the presentation layer of the system logic architecture is developed and implemented by Android programming. The business layer is implemented through Python network programming and C++ language. Among them, Python network programming is used for network communication, and the C++ package is responsible for decoding the speech sent by the presentation layer. The presentation layer is mainly responsible for recording voice clips and transmitting the voice clips to the remote server through the network. It does not contain any business logic related to the recognition function. The main work of the business logic layer is divided into three parts: receiving voice data, decoding, and feeding back the recognition result. The logic architecture adopted by this system better

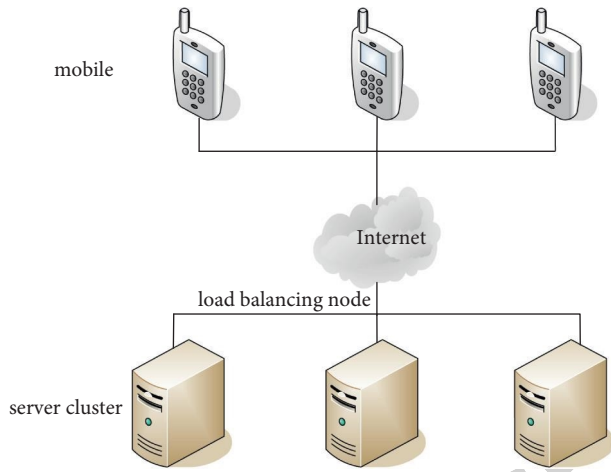


Figure 1: Physical architecture topology.

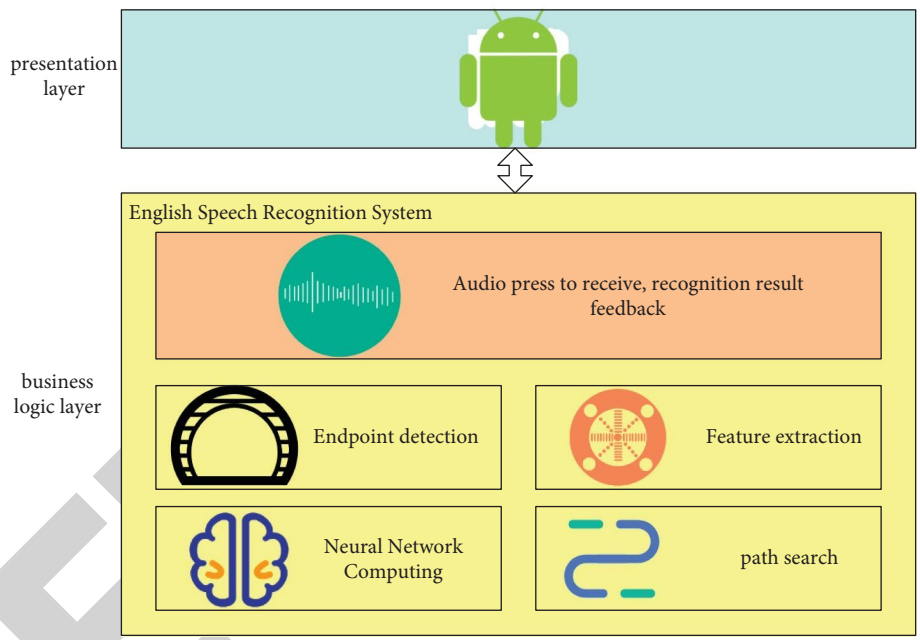


Figure 2: Logical structure.

reflects the principle of high cohesion and low coupling of software design, that is, the presentation layer. In addition to exchanging data, the business logic layer does not have any other business intersection. The advantage of this is strong reusability. Each layer implements its functions to facilitate cross-platform migration. Maintainability is good. When the system needs to be upgraded and maintained, the maintenance of one module will not affect the functions of other modules, reducing the workload of maintenance.

4.1.3. Design and Implementation of System Function Modules. In this section, the functions and processes of the entire system of system are first introduced. Then, the specific implementation of each module is described. The online speech recognition system mainly includes two parts, the first part is the client side, and the second part is the server side. The system flow chart is shown in Figure 3.

As shown in Figure 3, the client uses an Android-based mobile App client. The modules that need to be implemented are recording module, audio sending module, and recognition feedback receiving module. The modules that need to be implemented on the server side are mainly audio receiving module, endpoint detection module, feature extraction module, neural network calculation module, decoding path search module, and feedback recognition result.

Server-side modules are more complex than client-side modules. It mainly consists of six modules: audio reception, endpoint detection, feature extraction, neural network calculation, decoding path search, and recognition result feedback. The server-side functions are shown in Figure 4.

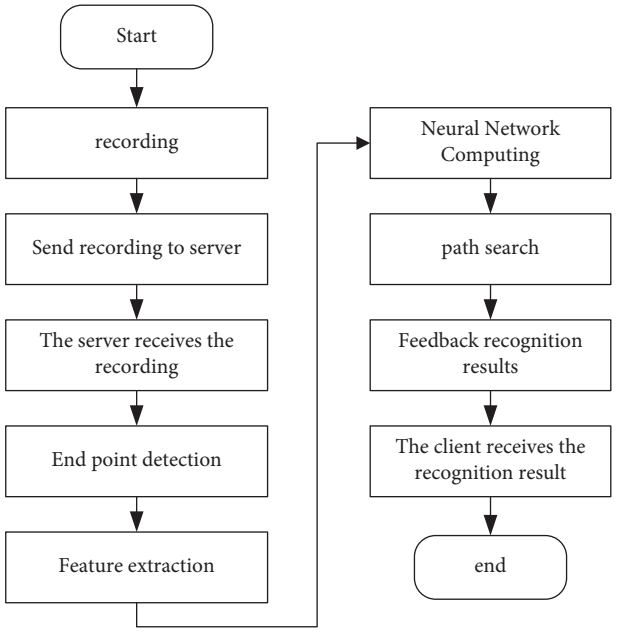


Figure 3: System flow chart.

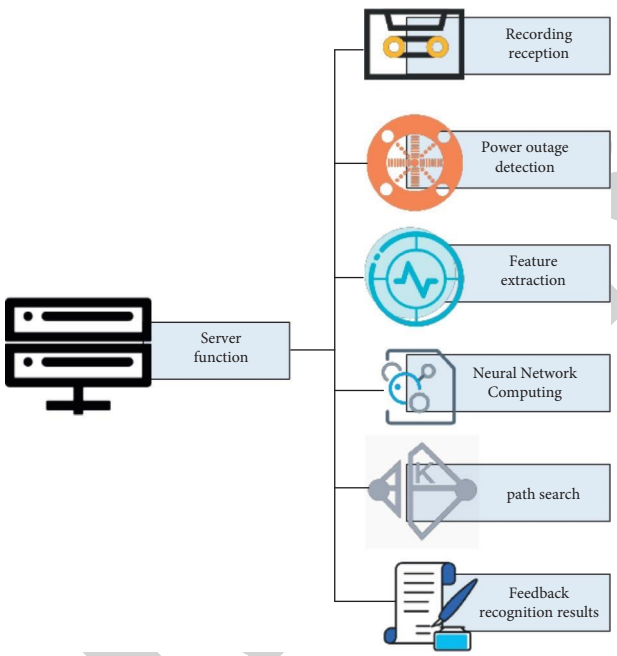


Figure 4: Server-side functions.

As shown in Figure 4, the server side is mainly implemented by Python network programming and C++ language. Among them, Python network programming is used for data reception and transmission and interaction with C++ modules. Since the speech recognition system requires high real-time performance, all modules related to recognition are implemented by C++ with high efficiency. The flow chart of the server side is shown in Figure 5.

As shown in Figure 5, the function of the server-side audio receiving and recognition result feedback module is simple, that is, conventional data sending and receiving,

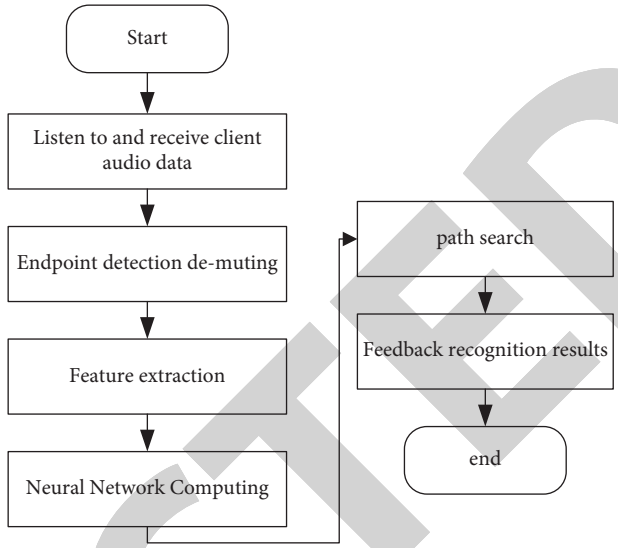


Figure 5: Server-side flowchart.

which is the same as the previous section. Considering the real-time nature of the speech recognition system, the communication protocol used on the server side is still the TCP protocol. The author uses the SocketServer class in Python network programming and uses the recv function in this class to accept audio data. When the identification result is sent, SocketServer provides two sending methods for data sending, send, and sendall. The send function is highly mobile. Each transmission may not necessarily send all the data, and usually, it will be repeated many times before the entire data are sent. To ensure the user experience of the system, the author uses the sendall function. This function will send all the data through one communication of TCP.

5. English Speech Optimization Teaching Recognition Modeling Simulation

5.1. Speech Recognition Experiment Based on Feature Fusion. To fully characterize the dynamic and static information of the speech signal, the new feature vector is put into the recognition network SVM for training and simulation. Based on the extracted features and its first-order difference, the TEOCC feature reflecting the signal energy change is added, and the obtained fusion feature set is used as the input of the recognition model SVM to further verify the superior performance of the fusion feature set in speech recognition. The experimental comparison results are shown in Figure 6.

As shown in Figure 6, COSR of Experiments 1 and 2, it can be seen that the FFPRLS_D features are more dominant in static features at 0dB20 words and 15dB20 words, which are 4.73% and 3.54% higher than FFPRLS features, respectively. From the point of view of the average recognition rate, the average recognition rates of the dynamic and static combined feature vectors have been improved to varying degrees, indicating that the combination of dynamic and static features can more effectively characterize the information of the speech signal, further confirming that the

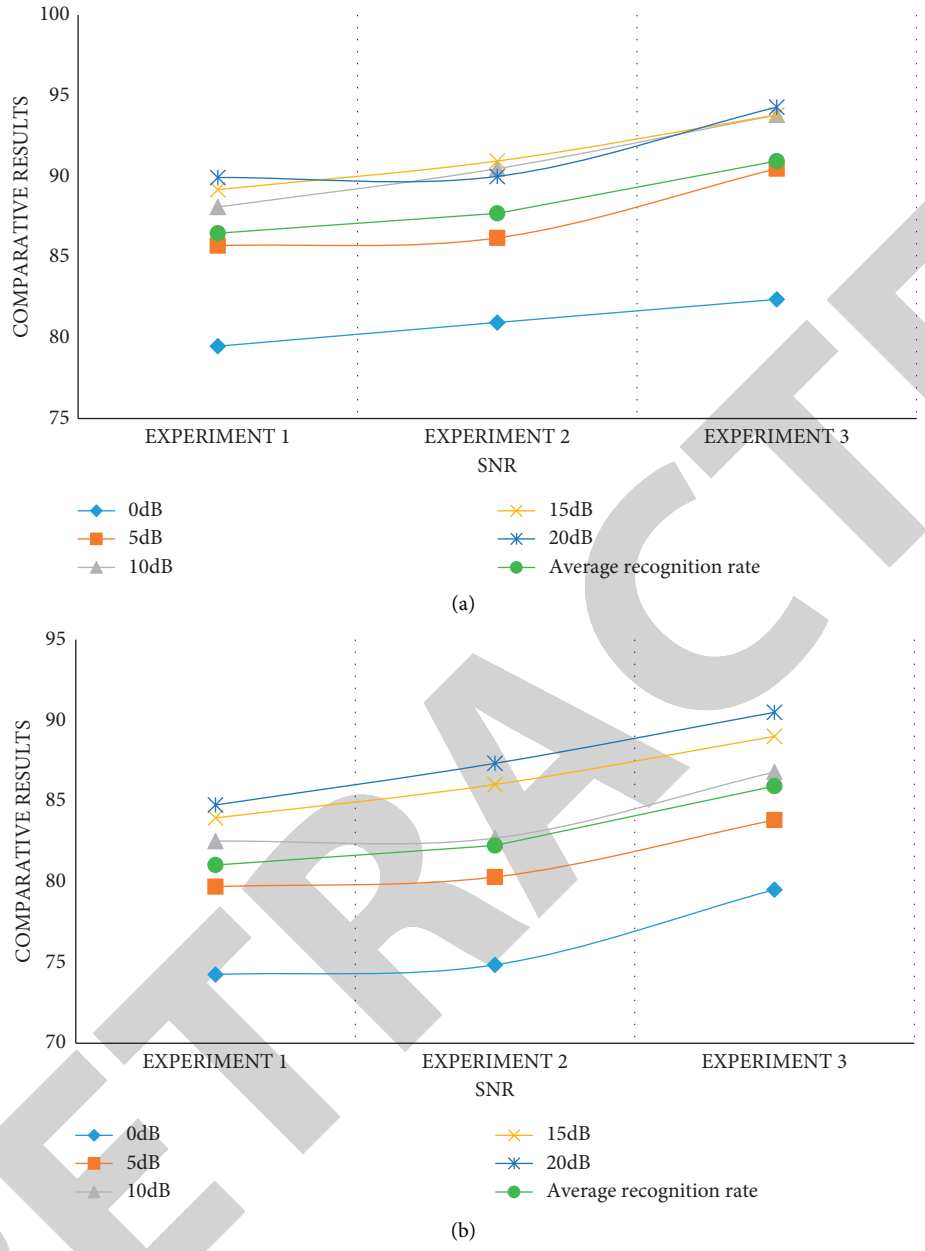


Figure 6: COSR results of FFPSS feature set. (a) COSR results of 10-word FFPSS feature set. (b) COSR results of 20-word FFFPS feature set.

combined feature vector is more effective than a single feature. The comparison of the speech recognition results of the FFPRLS feature set is shown in Figure 7.

As shown in Figure 7, comparing the recognition (COSR) results of Experiments 2 and 3, it can be seen that from the recognition rate under different SNR environments, after adding the TEOCC feature reflecting the energy change of the speech signal, compared with dynamic and static combined features, the recognition effect of fusion features has been further improved. The COSR results of FFPLMS feature set are shown in Figure 8.

As shown in Figure 8, from the perspective of the average recognition rate, compared with the dynamic and static combination of the three fusion features, the recognition rate in the case of 10 words is 2.69% higher on average, and the recognition rate in the case of 20 words is higher on average than 3.18%. The data show that the energy feature TEOCC contains the semantic information of the speech signal, which can be used as an auxiliary feature parameter to improve the performance of the speech recognition system.

5.2. Speech Recognition Experiment Based on Feature Optimization. To verify the effectiveness of the method proposed in this chapter to optimize the characteristic parameters based on PCA, simulation comparison experiments are carried out under different signal-to-noise ratios. Design the experimental scheme: The three dynamic and static combined features are optimized separately. The optimized feature set(OFS) is defined as PCA-Features1. Then

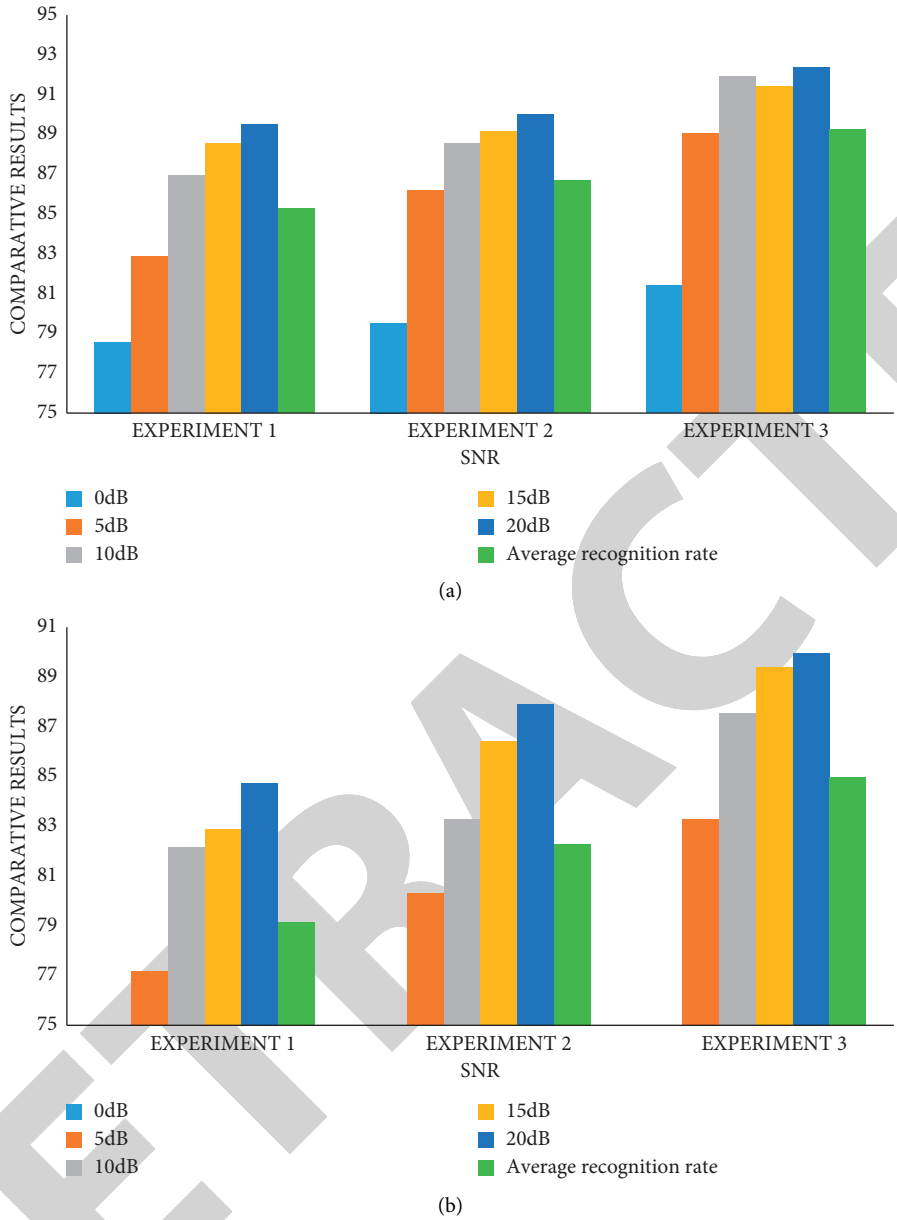


Figure 7: COSR results of FFPRLS feature set. (a) COSR results of 10 words. (b) COSR results of 20 words.

put the feature set into the recognition network SVM for training and simulation, and the obtained experimental comparison results are shown in Table 1;

As shown in Table 1, it can see that from the perspective of the average recognition rate, the average recognition rate of the OFS has been improved to varying degrees. From the perspective of recognition rates under different SNR environments, the OFS of FFPSS and FFPRLS have more prominent advantages under 0dB10 words. The recognition rates both increased by 7.15%. But at 0 dB and 10 words, the recognition rate of the OFS PCA-Features1 and FFPSS_D features is the same. Because PCA can retain the components containing important information in the feature parameters and remove some unnecessary components, it cannot completely retain the most effective information components.

The fused feature sets added to the energy feature TEOCC are respectively optimized for features. The OFS is defined as PCA-Features2. Then, the feature set is put into the recognition network SVM for training and simulation. The experimental comparison results obtained are shown in Table 2;

As shown in Table 2, comparing the results of Experiments 3 and 4, we can see that from the average recognition rate, the recognition effect of the OFS has been further improved compared with the fusion feature. From the point of view of the recognition rate under different SNR environments, the recognition rate of the OFS PCA-Features2 decreased by 0.47% compared with the FFPRLS_D + TEOCC feature set under 10dB10 words. Compared with the FFPLMS_D + TEOCC feature set, the

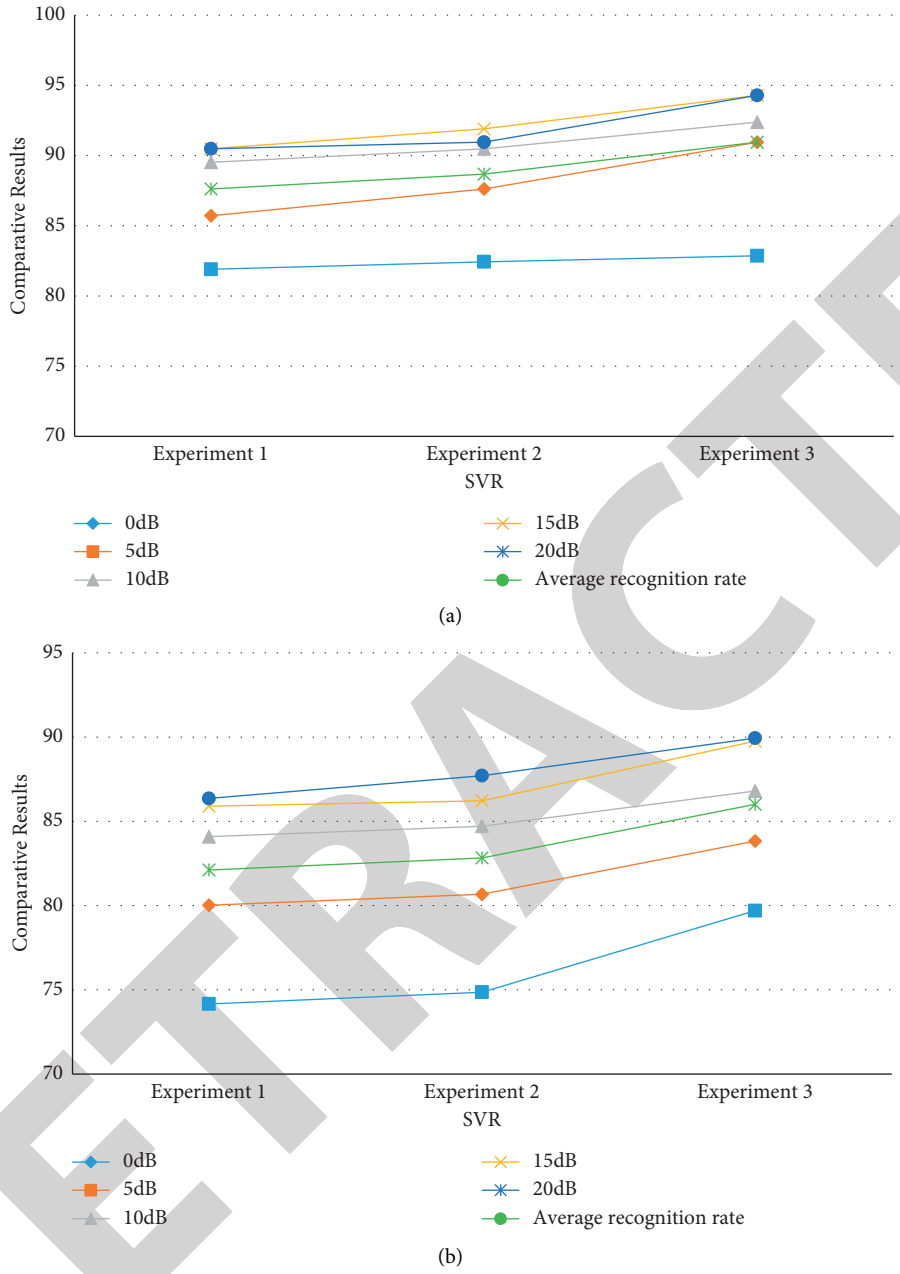


Figure 8: COSR results of FFPLMS feature set (%). (a) COSR results of 10 words. (b) COSR results of 20 words.

Table 1: COSR results before and after FFSS feature set optimization.

Vocabulary	Characteristic parameters	0 dB	5 dB	10 dB	15 dB	20 dB	Average recognition rate
10 words	FFPSS_D	80.95	86.19	90.48	90.95	90.00	87.71
	PCA-features1	88.10	89.52	90.48	92.38	92.38	90.57
	FFPSS_D + TEOCC	82.38	90.48	93.81	93.81	94.29	90.95
	PCA-features2	90.00	90.95	94.29	94.29	94.29	92.76
20 words	FFPSS_D	74.86	80.30	82.71	86.03	87.34	82.25
	PCA-features1	78.81	81.67	87.38	88.81	91.67	85.67
	FFPSS_D + TEOCC	79.52	83.83	86.80	89.01.	90.50	85.93
	PCA-features2	81.43	85.24	88.57	90.48	92.38	87.62

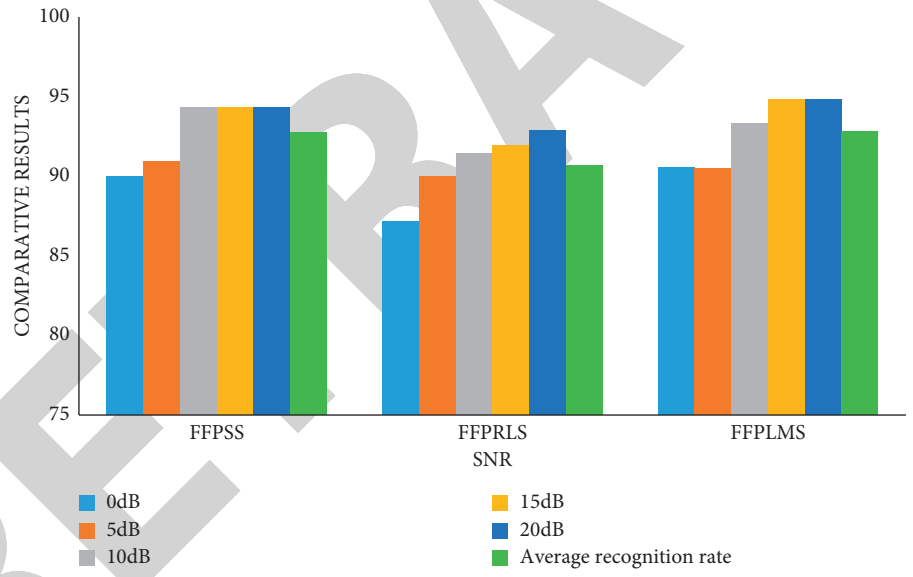
recognition rate under 5dB10 words also drops by 0.47%. This is also because PCA cannot completely remove redundancy and retain the most effective information, and its recognition effect becomes worse. The COSR results before and after FFPLMS feature set optimization is shown in Table 3.

Table 2: COSR results before and after FFPRLS feature set optimization.

Vocabulary	Characteristic parameters	0 dB	5 dB	10 dB	15 dB	20 dB	Average recognition rate
10 words	FFPSS_D	79.52	86.19	88.57	89.18	90.00	86.69
	PCA-features1	86.67	91.43	89.05	89.57	91.43	89.63
	FFPSS_D + TEOCC	81.43	89.05	91.90	91.43	92.38	89.24
	PCA-features2	87.14	90.00	91.43	91.95	92.86	90.68
20 words	FFPSS_D	73.56	80.30	83.27	86.41	87.90	82.29
	PCA-features1	78.57	82.14	87.38	87.95	89.76	85.16
	FFPSS_D + TEOCC	74.67	83.27	87.55	89.39	89.94	84.96
	PCA-features2	80.00	85.95	89.29	89.59	91.19	87.20

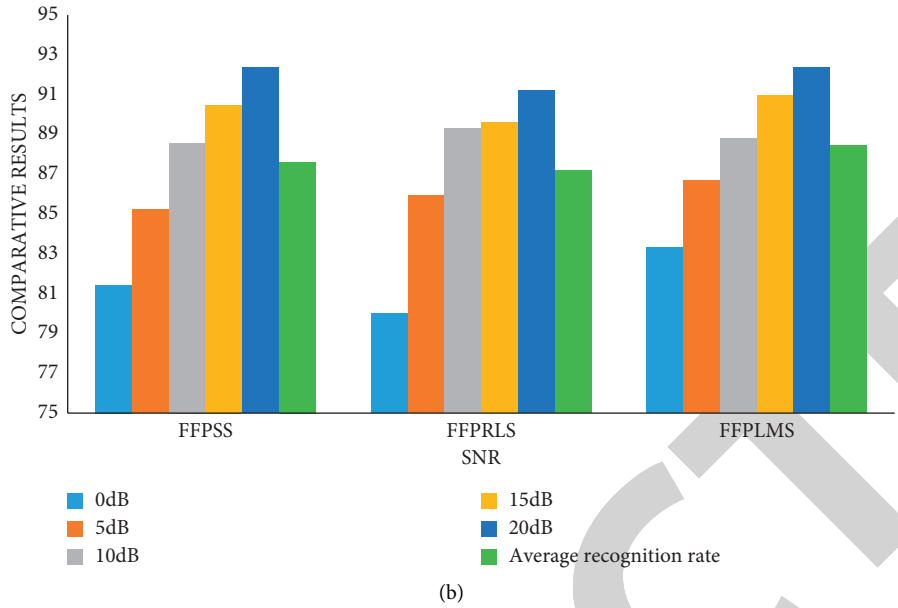
Table 3: COSR results before and after FFPLMS feature set optimization.

Vocabulary	Characteristic parameters	0 dB	5 dB	10 dB	15 dB	20 dB	Average recognition rate
10 words	FFPSS_D	82.43	87.62	90.48	91.90	90.95	88.68
	PCA-features1	88.57	89.52	93.33	91.95	92.38	91.15
	FFPSS_D + TEOCC	82.86	90.95	92.38	94.29	94.29	90.95
	PCA-features2	90.52	90.48	93.33	94.81	94.81	92.79
20 words	FFPSS_D	74.86	80.67	84.71	86.22	87.71	82.83
	PCA-features1	78.10	82.86	86.90	89.29	91.43	85.72
	FFPSS_D + TEOCC	79.70	83.83	86.80	89.76	89.94	86.01
	PCA-features2	83.33	86.67	88.81	90.95	92.38	88.43



(a)

Figure 9: Continued.



(b)

Figure 9: COSR results of three OFS. (a) COSR results of 10 words (%). (b) Comparison of 20-word speech recognition results.

The identification results of comprehensive comparison Experiments 1 and 2 and Experiments 3 and 4 are shown in Table 3: This chapter proposes to use PCA to perform feature selection on fused features to obtain an OFS, which reduces the feature dimension and obtains a higher recognition accuracy, thereby improving the recognition performance of the speech recognition system. It can be shown that the feature optimization method based on PCA can make up for the shortage of too many feature dimensions caused by feature fusion and can also retain most of the important information in the features.

To construct the optimal feature set, the advantages and disadvantages of the three OFS are compared as a whole, and the optimal feature set is further selected. The COSR results of the three OFS is shown in Figure 9.

As shown in Figure 9, it can be seen from the three sets of experimental results that: From the perspective of the recognition rate under different SNR environments, the FFPLMS OFS has the highest recognition rate when the SNR is 0 dB. From the average recognition rate, whether it is 10 words or 20 words, the FFPLMS OFS shows a good recognition effect. After a comprehensive analysis, the recognition performance of the above three feature sets is ranked as follows: the FFPLMS OFS is the best, the FFPSS OFS is the second, and the FFPRLS OFS is the last. The optimal feature set thus obtained is the FFPLMS OFS. The feature set can better describe the characteristics of the speech signal.

6. Conclusions

As the speech signal is not smooth in time, the Fourier transform of the speech signal has certain limitations. Therefore, this paper adopts the acoustic transform to analyze the speech signal, applies the acoustic CFCC function to the speech recognition system, and proposes a new function to improve the nonlinear transformation process of the CFCC function. It improves the accuracy of the neural network while increasing the training speed and decoding speed of the model and proposes a new model on how to build a complete English speech optimized teaching recognition system using an intelligent edge detection algorithm. The whole system includes the mobile side and the server side and implements the most basic business logic of the speech recognition system. The auditory model includes multiple stages of nonlinear transformations. This paper only uses nonlinear power functions to simulate the auditory characteristics of the human ear, which is difficult to quantitatively analyze. Therefore, further study of auditory models is required.

Data Availability

The data that support the findings of this study are available from the corresponding author upon reasonable request.

Conflicts of Interest

The author(s) declare no potential conflicts of interest with respect to the research, author-ship, and/or publication of this article.

References

[1] V. Subba Ramaiah and R. Rajeswara Rao, "A novel approach for speaker diarization system using TMFCC parameterization and Lion optimization," *Journal of Central South University*, vol. 24, no. 11, pp. 2649–2663, 2017.

[2] S. Misirov, "The peculiarities of teaching English pronunciation in elementary classes (grades)," *Scientific Bulletin of Namangan State University*, vol. 1, no. 2, p. 63, 2019.

[3] H. Vančová and skTrnava University, "Current issues in pronunciation teaching to non-native learners of English," *Journal of Language and Cultural Education*, vol. 7, no. 2, pp. 140–155, 2019.

[4] A. Gashaw, "Rhythm in Ethiopian English: implications for the teaching of English prosody," *International Journal of Education and Literacy Studies*, vol. 5, no. 1, p. 13, 2017.

[5] F. Chika, E. Sayoko, and A.-Y. Reiko, "An evaluation of English pronunciation of Japanese EFL learners using multiple metrics," *Journal of the Phonetic Society of Japan*, vol. 22, no. 2, pp. 39–43, 2018.

[6] O. Kolesnikova, "Comparative analysis of American English and Mexican Spanish consonants for computer assisted pronunciation training," *Revista Signos*, vol. 50, no. 94, pp. 195–216, 2017.

[7] K.-I. Han, H.-J. Park, and K.-M. Lee, "Speech recognition and lip shape feature extraction for English vowel pronunciation of the hearing - impaired based on SVM technique," *Journal of Rehabilitation Welfare Engineering & Assistive Technology*, vol. 11, no. 3, pp. 247–252, 2017.

[8] E. Zajdler and zajdler@uj.edu.plJagiellonian University, "Speech shadowing as a teaching technique in the CFL classroom," *Lingua Posnaniensis*, vol. 62, no. 1, pp. 77–88, 2020.

[9] J. Fouz-González, "Podcast-based pronunciation training: enhancing FL learners' perception and production of fossilised segmental features," *ReCALL*, vol. 31, no. 2, pp. 150–169, 2019.

[10] H. Wei, Y. Long, and H. Mao, "Improvements on self-adaptive voice activity detector for telephone data," *International Journal of Speech Technology*, vol. 19, no. 3, pp. 623–630, 2016.

[11] H. Wei and N. Kehtarnavaz, "Determining number of speakers from single microphone speech signals by multi-label convolutional neural network," in *Proceedings of the IECON 2018 - 44th Annual Conference of the IEEE Industrial Electronics Society*, October, 2018.

[12] C. C. R. Yang and Y. Chen, "Implementing the flipped classroom approach in primary English classrooms in China," *Education and Information Technologies*, vol. 25, no. 2, pp. 1217–1235, 2019.

[13] G. Kartal and S. Korucu-Kis, "The use of Twitter and Youglish for the learning and retention of commonly mispronounced English words," *Education and Information Technologies*, vol. 25, no. 1, pp. 193–221, 2020.

[14] T. Isaacs and L. Harding, "Pronunciation assessment," *Language Teaching*, vol. 50, no. 3, pp. 347–366, 2017.

[15] R. Panahi and I. Gholampour, "Accurate detection and recognition of dirty vehicle plate numbers for high-speed applications," *IEEE Transactions on Intelligent Transportation Systems*, vol. 18, no. 4, pp. 767–779, 2017.

[16] T.-Y. Kim, H. Ko, S.-H. Kim, and H.-D. Kim, "Modeling of recommendation system based on emotional information and collaborative filtering," *Sensors*, vol. 21, no. 6, p. 1997, 2021.

[17] J. Linke, G. J. McDermid, A. D. Pape et al., "The influence of patch-delineation mismatches on multi-temporal landscape pattern analysis," *Landscape Ecology*, vol. 24, no. 2, pp. 157–170, 2009.

[18] Y. Zhang, D. Lefebvre, and Q. Li, "Automatic detection of defects in tire radiographic images," *IEEE Transactions on Automation Science and Engineering*, vol. 14, no. 3, pp. 1378–1386, 2017.

[19] S. Craciun, R. Kirchgessner, A. D. George, H. Lam, and J. C. Principe, "A real-time, power-efficient architecture for mean-shift image segmentation," *Journal of Real-Time Image Processing*, vol. 14, no. 2, pp. 379–394, 2018.

[20] A. J. Magana and T. De Jong, "Modeling and simulation practices in engineering education," *Computer Applications in Engineering Education*, vol. 26, no. 4, pp. 731–738, 2018.

[21] N. Borhan and E. Zakaria, "Structural equation modeling assessing relationship between mathematics beliefs, teachers' attitudes and teaching practices among novice teachers in Malaysia," *AIP Conference Proceedings*, vol. 1847, no. 1, pp. 1–7, 2017.

[22] S. Xinhang, "Shuqiang. Multi-scale multi-feature context modeling for scene recognition in the semantic manifold," *IEEE Transactions on Image Processing*, vol. 26, no. 6, pp. 2721–2735, 2017.

[23] L. Morell, T. Collier, P. Black, and M. Wilson, "A construct-modeling approach to develop a learning progression of how students understand the structure of matter," *Journal of Research in Science Teaching*, vol. 54, no. 8, pp. 1024–1048, 2017.

[24] G. Li, F. Liu, A. Sharma et al., "Research on the natural language recognition method based on cluster Analysis using neural network," *Mathematical Problems in Engineering*, vol. 2021, Article ID 9982305, 13 pages, 2021.

[25] F. Meng, S. Yang, J. Wang, L. Xia, and H. Liu, "Creating knowledge graph of electric power equipment faults based on BERT–BiLSTM–CRF model," *J. Electr. Eng. Technol.*, 2022.

[26] J. Cao, Y. Xiang, Y. Zhang, Z. Qi, X. Chen, and Y. Zheng, "CONNER: a cascade count and measurement extraction tool for scientific discourse," in *Proceedings of the 15th International Workshop on Semantic Evaluation. (SemEval-2021*, pp. 1239–1244, Bangkok, Thailand, August, 2021.