Hindawi
Wireless Communications and Mobile Computing
Volume 2022, Article ID 2882964, 11 pages
https://doi.org/10.1155/2022/2882964

Research Article

Research on English Vocabulary and Speech Corpus Recognition Based on Deep Learning

Wang Zhen

Department of Public Education, Inner Mongola Techical College Of Construction, Hohhot 010070, China

Correspondence should be addressed to Wang Zhen; b20160904105@stu.ccsu.edu.cn

Received 8 June 2022; Revised 14 August 2022; Accepted 22 August 2022; Published 19 September 2022

Academic Editor: Jun Ye

In order to investigate how to recognize English words and speech corpus, an English vocabulary and English speech recognition model based on deep learning algorithm was proposed. Through recommending key technical problems and solutions based on deep learning algorithm, how to realize the recognition of English vocabulary and speech corpus was investigated. In the research, the accuracy of the method on the English vocabulary and speech corpus recognition based on the deep learning algorithm increased 79% over the previous methods. Combined with the principle of the deep automatic encoder and deep learning algorithm, the research emphasis was on the effects of speech recognition framework for speech corpus. The speech recognition research based on the theory of deep learning not only had a theoretical guidance meaning but also had the use value in the practical application.

1. Introduction

Due to the complex changes in speech pronunciation, the large amount of data of speech signals, the high dimension of speech feature parameters, and the large amount of computation for speech recognition and evaluation, high-demand software and hardware resources and algorithms are required for large-scale speech signal processing. However, the traditional speech recognition algorithm dynamic time warping algorithm, hidden Markov model, and artificial neural network have their own advantages and disadvantages, they have encountered unprecedented bottlenecks, and it is difficult to further improve their accuracy and speed. In recent years, with the development of deep learning research in the field of machine learning and the accumulation of big data corpus, speech recognition and evaluation technology has developed rapidly. Language is an essential element in people's daily communication. Speech is also an essential means of information exchange in people's daily life and work. Clear speech expression can further clarify the expression of information and further simplify the main idea that people want to express. In addition, a large piece of relatively complex information can be decomposed into different parts to help people better communicate and analyze [1]. But with the development of information technology and Internet technology, the emergence and development of speech recognition has gradually changed the people's living habits. By using intelligent terminal, computer, intelligent wear equipment, speech recognition can be realized. Speech recognition technology has become one of the technical means of language communication in today's society. How to better apply this technology to assist people's communication is the focus of research [2]. Therefore, combining the principle of deep autoencoder and deep learning algorithm, an English vocabulary and English speech recognition model based on deep learning algorithm is proposed, which focuses on the influence of speech recognition framework on speech corpus.

Speech recognition technology was mainly divided into the following stages. In 1950s, Audrey in AT&T Bell Laboratory was the prototype of speech recognition. In late 1960s and early 1970s, it has a significant progress [3]. In late 1980s, it has a breakthrough. In early 1990s, many large companies launched their own speech recognition apps. The Audrey system, first developed at AT&T Bell Labs in the 1950s, was the first speech recognition system capable of

recognizing 10 English digits. However, substantial progress was made in the late 1960s and early 1970s [4]. The main reason was the introduction of linear predictive coding plane (LPC) technology and dynamic time warping (DTW) technology, which could effectively solve the problem of feature extraction and unequal length matching of speech signal. Speaking skills at that time were generally based on the principle of template matching. And the name is limited to identifying the difference between special people and small words. The identification of specific population segregation based on cepstral prediction and DTW techniques has been observed. At the same time, vector quantization (VQ) and hidden Markov model (HMM) theory were proposed [5].

In the late 1980s, lab speech recognition research finally had a breakthrough. For the first time in the lab, the barriers of large vocabulary, continuous speech, and nonspecific people were by combining all three characteristics in one system, typically Carnegie Mellon University's Sphinx system. It was the first high-performance nonspecific large vocabulary continuous speech recognition system. During this period, speech cognition research was further understood, characterized by the use of HMM models and neural network devices in speech cognition [6]. Zhang et al. at AT&T Bell Labs evaluated the HMM model for a wide range of applications. They engineered the original difficult HMM pure mathematical model, so that more researchers could understand it and make statistical methods that will become the mainstream of speech recognition technology. Statistical methods shifted researchers' attention from micro to macro. They no longer deliberately pursued refinement of speech features but constructed the best speech recognition system more from the overall average (statistical) perspective [7]. The research on speech recognition in China started in the 1950s. The Institute of Acoustics of Chinese Academy of Sciences began to conduct speech research. The real beginning of speech recognition in China should be the generation of RTSRS(01), a speech recognition system which used bandpass filter bank parameters and realized by Institute of Acoustics of Chinese Academy of Sciences in 1978. In the 1980s, professors from Tsinghua University proposed an implicit Markov model based on segment state distribution, which effectively solved the pruning problem of language identification in multilingual continuous recognition system [8]. In 2004, some scholars used HMM and GMM to score Chinese pronunciation and tone, respectively. Downhill Simplex Search optimized subsystem parameters in order to achieve the same scoring standard consistent with Chinese experts. Fluent application systems included FLUENCY of Language Technology Research Institute of Carnegie Mellon University and School of Information of Kyoto University in Japan. Some institutions in China, such as PLASER at Hong Kong University of Science and Technology, Department of Electronic Engineering at Tsinghua University, Department of Computer Science of Harbin Institute of Technology, and the Department of Computer Science of Harbin Institute of Technology, have also made some significant progress in these researches. However, most researches in China were to assist the learning of Chinese pronunciation [9].

2. Methods

2.1. Key Technologies of Deep Learning

2.1.1. Energy Probability Model. Introducing RBM into network modeling is a breakthrough with theoretical guiding significance for deep neural networks [10]. Using RBM as an energy model, it is possible to model arbitrarily distributed data. The Boltzmann machine is a large class of neural network models, but the most commonly used one in practice is the RBM. The RBM itself is simple, just a two-layer neural network, so it is not strictly considered as a category of deep learning. When the minimum energy of the overall network is calculated iteratively, it means that the system is in steady state at this time and the network parameters we require are also the network parameters at this time.

2.1.2. Pretraining Layer by Layer. In the past, neural networks determined the initial value through random initialization, at which time the random option value was required. It was often inconsistent with the actual situation, so the final effect was not ideal [11]. With RBM, the model is built in the middle of the two adjacent layers, and the training is carried out layer by layer from bottom to top. After several iterations, RBM enters a relatively stable state. Each neuron in the visible layer is connected to all the neurons in the hidden layer, but there is no connection between the neurons in the same layer, and all the neurons have only two output states. In this case, the hidden layer and the visible layer are equivalent to the same features in more than one space in different expressions, so as to determine the initial value consistent with the weight of the actual situation [12].

2.1.3. Network Parallel Training. Considering that there are several hidden layers in the deep neural network, each of which has more than 1000 nodes, the number of relevant parameters is likely to exceed 1 million. In such a large-scale network, the time of data training will be greatly extended without parallel processing. The BP neural network is mainly composed of the input layer, the hidden layer, and the output layer. The number of nodes in the input and output layer is fixed. Whether it is a regression or a classification task, choosing the appropriate number of layers and the number of hidden layer nodes will affect the performance of the neural network to a large extent. Parallel processing can be completed by hardware or software. The hardware method requires the support of GPU or distributed computing cluster. The software method means that the parameters of data subset are updated by multithreading and the updated results are unified at an appropriate time to complete the parallel training of the network [13].

2.2. Encoder Category Based on Depth Theory

2.2.1. Deep Autoencoder. The input required by the deep autoencoder is the original data feature. And the middle

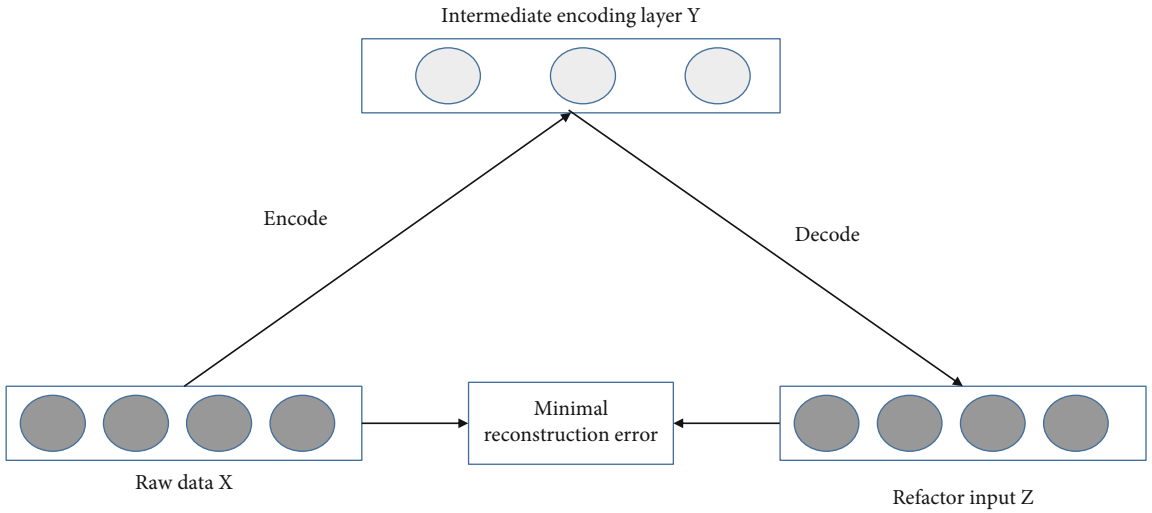

Figure 1: Autoencoder model.

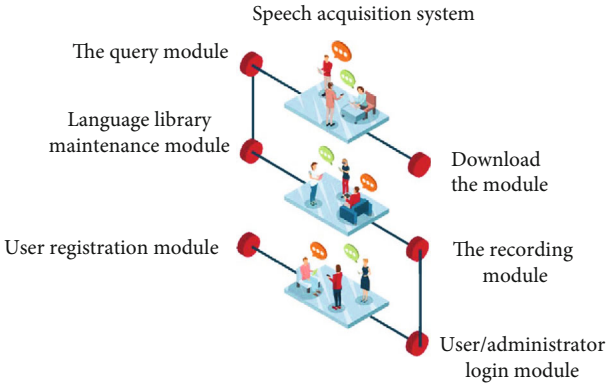

Figure 2: Speech acquisition system module.

layer encoding feature is obtained by different hidden layer encoding and the original input is reconstructed according to the decoding. Autoencoder is a kind of neural network, whose basic idea is to directly use one layer or more layer of neural network to map the input data and get the output vector. The model is shown in Figure 1. Network parameter adjustment is mainly aimed at minimizing the mean square error between original input and reconstructed input. The calculation method of loss function is shown in

$$J(W,b)=\left[\frac{1}{m}\sum_{i=1}^{m}J\left(W,b,x^{(i)},y^{(i)}\right)\right]+\frac{\lambda}{2}\sum_{l=1}^{n_i-1}\sum_{i=1}^{s_l}\sum_{j=1}^{s_{l+1}}\left(W_{ji}^{(l)}\right)^2, \tag{1}$$

$$J(W,b)=\left[\frac{1}{m}\sum_{1}^{m}\left(\frac{1}{2}\left\|x^{(i)}-h_{W,b}x^{(i)}\right\|^2\right)\right]+\frac{\lambda}{2}\sum_{l=1}^{n_i-1}\sum_{i=1}^{s_l}\sum_{j=1}^{s_{l+1}}\left(W_{ji}^{(l)}\right)^2. \tag{2}$$

In formula (1) and formula (2), the first term represents average reconstruction error. The second term represents regularization constraint term, aiming to prevent overfitting [14]. m represents the amount of training data. W and b are parameters of the encoder. $x^{(i)}$ and $y^{(i)}$ represent the original input and reconstruction input in turn, and their relationship is shown in

$$y^{(i)}=h_{W,b}\left(x^{(i)}\right). \tag{3}$$

2.2.2. Denoising Autoencoder. The training data required for this encoder is random noise that is superimposed on the raw data before providing it to the network (adding random noise to input layer nodes or according to some probability to make some input layer nodes 0). After the coding module is used to obtain the coding representation of the middle layer, the original data is reconstructed on the output layer to obtain more prominent features in robustness. The original data layer (the data in it is the raw data, without any processing) is the original json format data, because the original data has two kinds of data: start log and event log.

2.2.3. Sparse Autoencoder. Sparse autoencoder, another important extension model of autoencoder, also has good feature extraction performance. Sparse means that the hidden layer node has a high probability of 0 and sparse autoencoder is an unsupervised machine learning algorithm that constantly adjusts the parameters of the autoencoder by calculating the error between the autoencoding output and the original input to finally train the model. Autoencoders can be used to compress the input information and extract useful input features, and its non-0 time is relatively short (there is a long distance between it and 0; that is, it is in active state) [15]. Research on the visual perception system of human brain shows that the distribution of visual cortex cells in V1 region is sparse after the human brain receives natural image signals, even though only a few of them are activated at the same time. The output state of the hidden layer of the network is limited, so that the nodes of the hidden layer enter the sparse state, and the average output of the nodes of the hidden layer is equal to 0. In this way, the proportion of active nodes is relatively small, and the homogeneity of

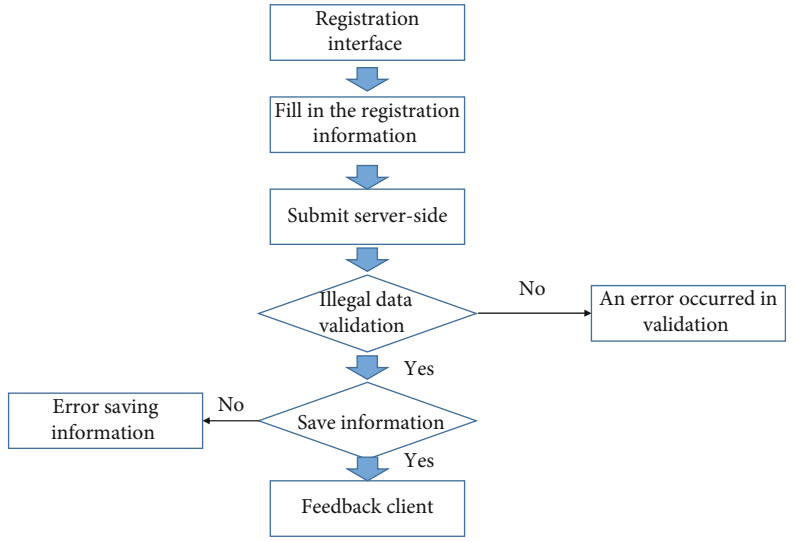

Figure 3: User registration process.

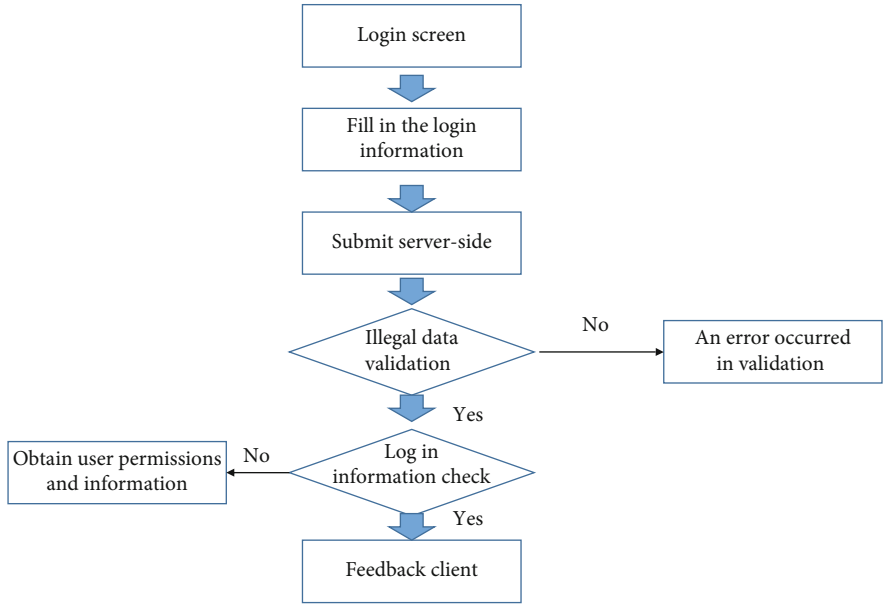

Figure 4: Login module process.

the characteristics of the nodes of the hidden layer will not occur [16]. The loss function of sparse autoencoder is shown in

$$J_{\text{sparse}}(W,b) = J(W,b) + \beta \sum_{j=1}^{s_2} \text{KL}\left(\rho \| \bar{\rho}_j\right). \tag{4}$$

The first term of formula (4), which is the same as formula (1), represents the size of reconstruction error. The second term is KL distance, representing the gap between the expected sparsity and the actual value, which can be calculated by the following expression, as shown in

$$\text{KL}\left(\rho \| \bar{\rho}_j\right) = \rho \log \frac{\rho}{\bar{\rho}_j} + (1-\rho) \log \frac{1-\rho}{1-\bar{\rho}_j}. \tag{5}$$

$\bar{\rho}_j$ represents the average output value of nodes at the hidden layer, which satisfies

$$\bar{\rho}_j = \frac{1}{m} \sum_{i=1}^{m} \left[a_j^{(2)}\left(x^{(i)}\right)\right]. \tag{6}$$

2.3. System Framework Based on Deep Autoencoder

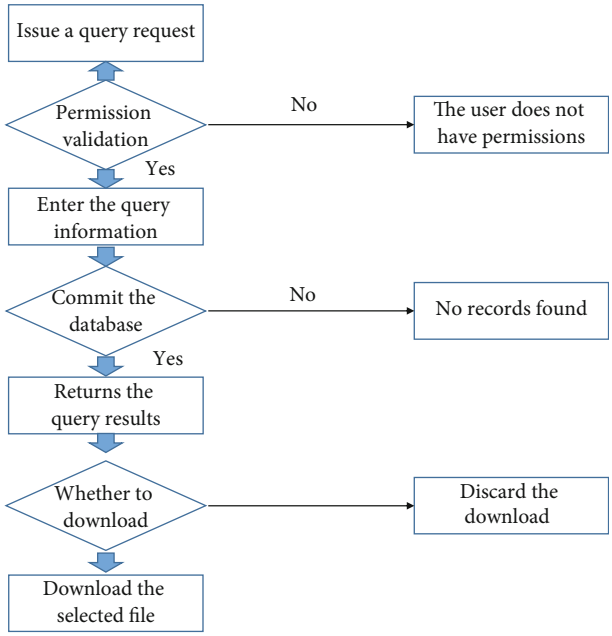

Figure 5: Query and download module flow.

2.3.1. Experimental Corpus. The original data required in the experiment are all from TIMIT speech data set. The full name of TIMIT is The DARPA TIMIT Acoustic-Speech Continuous Speech Corpus, which is collected and constructed by Texas Instruments, Massachusetts Institute of Technology, and Stanford Research Institute. There are 6,300 sentences sampled at 16 kHz from eight different locations in the United States, and all sentences are manually segmented and labeled.

2.3.2. Feature Preprocessing. In the process of extracting high-level features, deep neural networks generally need to receive acoustic features such as MFCC and Fbank. Because of the copronunciation phenomenon, it is necessary to extract digital features from images (or texts) for use by various models. Sometimes, you need to extract numerical features from images (or text) for use by various models. Deep learning models can be used not only for classification regression but also for extract features. The trained model is usually used to input pictures and output as extracted feature vectors. It is generally necessary to expand the short-term features to obtain the superframe features carrying context information. Original feature extraction is as follows: according to the parameters of frame length 20 ms and frame shift 10 ms, the 39-dimensional MFCC features (12-dimensional output + 1-dimensional logarithmic energy and their first- and second-order differences) are extracted from the original speech through the HCopy file provided by HTK. A voice sample in the data for detailed description is selected. First, two text files should be created in the same root directory, named YL. conf and YL. scp, respectively. The former mainly writes parameters for MFCC extraction, and the latter is the path of sample files and generated files. The yangli.mfc file can be obtained in the same directory after the extraction is successful. Since the file format cannot be directly viewed, the HList tool can be used to convert it to a txt file.

Data preprocessing is as follows: 5 frames are added before and after the features obtained in the previous step to obtain 11 consecutive superframe features. Then, the cepstrum mean variance is normalized. The processed features are input through the visibility layer as training samples of the network model [17]. In the process of normalization of each dimension of superframe feature, the two points cannot be ignored. First, normalization can reduce the influence caused by feature difference between channel and individual. Second, Gauss-Bernoulli RBM model is selected in the process of modeling the input layer and the first hidden layer whose node states conform to Gaussian distribution. At this time, the energy function is shown in

$$E(v,h)=\sum_{i\in V}\frac{(v_i-a_i)^2}{2\sigma_i^2}+b^Th-\sum_{i\in V,j\in H}\frac{v_i}{\sigma_i}h_iw_{ij}. \quad (7)$$

After CMVN processing, input data distribution in formula (7) satisfies

$$\begin{cases} a_i=0, \\ \sigma_i=1. \end{cases} \quad (8)$$

The energy function is equivalent to

$$E(v,h)=\sum_{i\in V}\frac{v_i^2}{2}-b^Th-\sum_{i\in V,j\in H}v_ih_iw_{ij}. \quad (9)$$

2.3.3. Autoencoder Structure. The structure of the encoder includes the number of hidden layers, the number of nodes contained in each hidden layer, and the node type of each hidden layer [18]. After many experiments, the deep autoencoder used in the study consists of seven layers, including an input layer, an output layer, and five hidden layers, and the number of nodes in each layer is $490\times720\times720\times50\times720\times720\times490$ [19].

2.3.4. Network Training Algorithm

(1) Gauss-Bernoulli RBM Training. Because the network input has the speech cepstrum feature, the value is between $[-\infty, +\infty]$, which is obviously different from the black and white image signal. Gauss-Bernoulli RBM is often selected as the input layer and the first hidden layer to build the model. In practice, data preprocessing is needed to normalize the input feature mean and variance. The first several layers of the model are mainly divided into visible layer (490 Gauss nodes) and hidden layer (720 Bernoulli nodes). Here is the training algorithm.

(1) Given a sample v of training data, the activation probability of hidden layer node hj can be expressed as shown in

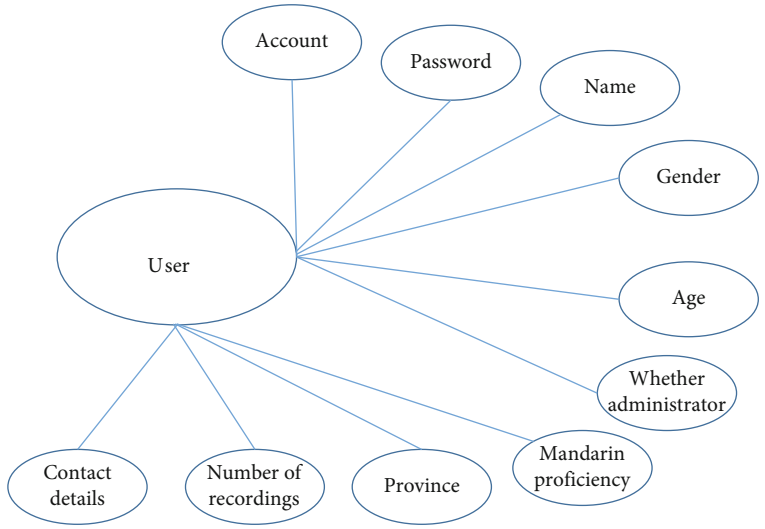

Figure 6: User entity attribute diagram.

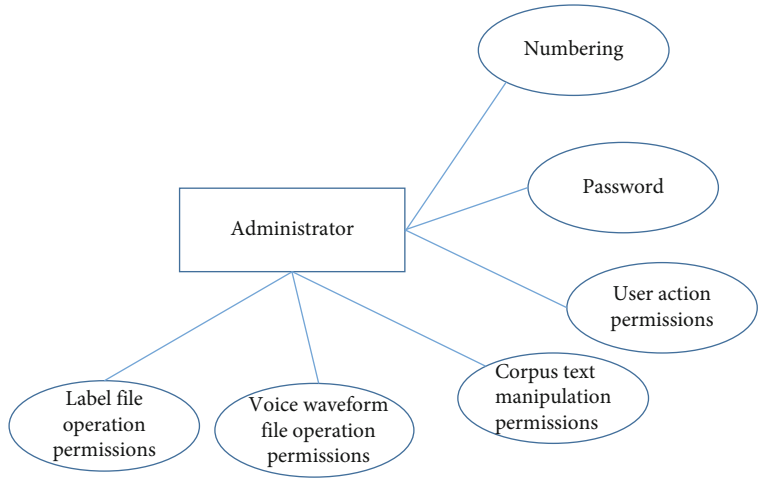

Figure 7: Administrator entity diagram.

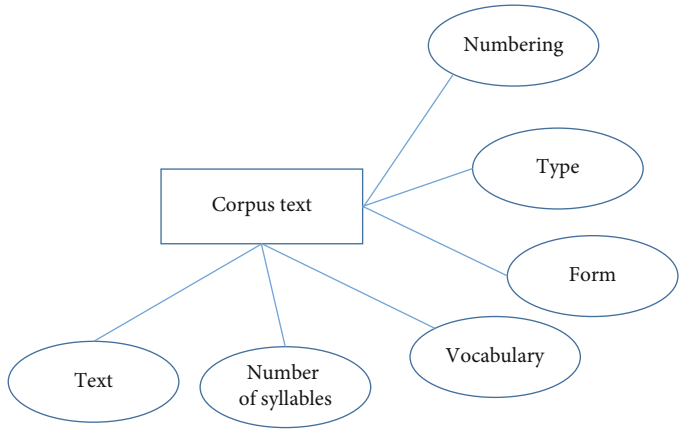

Figure 8: Corpus text entity diagram.

$$p(h_j = 1|v) = \sigma\left(b_j + \sum_{i\in \text{vis}} v_i w_{ij}\right) \tag{10}$$

(2) Randomize the hidden layer node values obtained in (1) to generate 0 and 1 activation states, and deduce the visible layer input v' according to the hidden layer node states. For the linear visible layer element, its reconstruction formula is expressed as

$$v' = N\left(b_i + \sum_{j\in hid} h_j w_{ij}, 1\right) \tag{11}$$

(3) The reconstructed visible layer state value v is used as the input of RBM structure. The hidden layer probability h is calculated again according to step (1)

(4) Update weight parameters according to formula (12), where <. > is the average value of all samples in each small batch and ε is the learning rate, as shown in

$$\Delta w_{ij} = \varepsilon\left(<v_i h_j> - <v_i' h_j'>\right) \tag{12}$$

The initialization parameters of the model are as follows: the weight of the connection is set to a small value and the node bias is set to 0. When each size is done, there are 256 minibatch models. Degrees are 0.01. The activation probability value of each node of the last training hidden layer h1 is retained as the input data of visible layer of RBM in the upper-middle layer of superposition structure [20].

(2) Bernoulli-Bernoulli RBM Training. The output value of the hidden layer of the first Bernoulli-Bernoulli RBM model is directly defined as the input value of the visible layer of the next RBM, and then, the connection weight between $h1$ and $h2$ of the hidden layer is continued to be trained. Compared with Gauss-Bernoulli RBM, the training method is basically similar, but the visible layer nodes obey Bernoulli distribution. Here, the hidden layer state is used to reconstruct the visible layer, and the basic formula is shown in

$$p\left(v_i' = 1|h\right) = \sigma\left(b_i + \sum_{i\in \text{hid}} h_i w_{ij}\right). \tag{13}$$

(3) Network Parameter Tuning. After the initial model is pretrained, network parameters need to be adjusted, usually through backpropagation (BP) [21]. The sample overall loss function can be written as shown in

$$J(W, b) = \left[\frac{1}{m}\sum_{i=1}^{m} J\left(W, b, x^{(i)}, y^{(i)}\right)\right] + \frac{\lambda}{2}\sum_{l=1}^{n_i-1}\sum_{i=1}^{s_l}\sum_{j=1}^{s_{l+1}}\left(W_{ji}^{(l)}\right)^2, \tag{14}$$

$$J(W, b) = \left[\frac{1}{m}\sum_{i=1}^{m}\left(\frac{1}{2}\left\|h_{W,b}\left(x^{(i)}\right) - y^{(i)}\right\|^2\right)\right] + \frac{\lambda}{2}\sum_{l=1}^{n_i-1}\sum_{i=1}^{s_l}\sum_{j=1}^{s_{l+1}}\left(W_{ji}^{(l)}\right)^2. \tag{15}$$

The first term of formula (14) is the mean square deviation term, which reflects the degree of difference between reconstruction and original input features. The second term is added to avoid the overfitting problem, which is the so-called regularization term. The contribution of the first and second terms to the loss function can be balanced by increasing the weight attenuation parameter λ. $h_{W,b}(x^{(i)})$ represents the reconstruction result obtained through the process of coding and decoding the sample $x^{(i)}$ through the network. The present invention provides a research on the recognition of English vocabulary and speech corpus based on a deep learning algorithm, which comprehensively evaluates the English pronunciation quality of the preset object through the two different aspects of the English pronunciation and the English vocabulary, so that it can comprehensively evaluate the English pronunciation quality of the preset object. It can accurately evaluate the actual English pronunciation accuracy and standard degree of the preset object and give an objective and reliable pronunciation quality evaluation score accordingly, so as to effectively improve the English pronunciation quality and improve the experience of learning English.

3. Results and Analysis

The design of the overall structure of the system is to reasonably divide the whole system into various functional modules, so as to correctly handle the relationship between and within modules, as well as the data connection between them, and then to define the internal structure of each module [22].

3.1. Structural Design of the System. In the system, C/S system architecture is used, including five function modules, namely, user registration module, user administrator login module, recording module, database maintenance module, and query and download module (see Figure 2).

3.2. Detailed Design

3.2.1. System Module

(1) User Registration Module. The registration module realizes the user registration function and corresponds the recording information with the account, which is convenient for users to query and use in the future. Account number, password, age, gender, Mandarin proficiency, and native place will be written into the user information form as

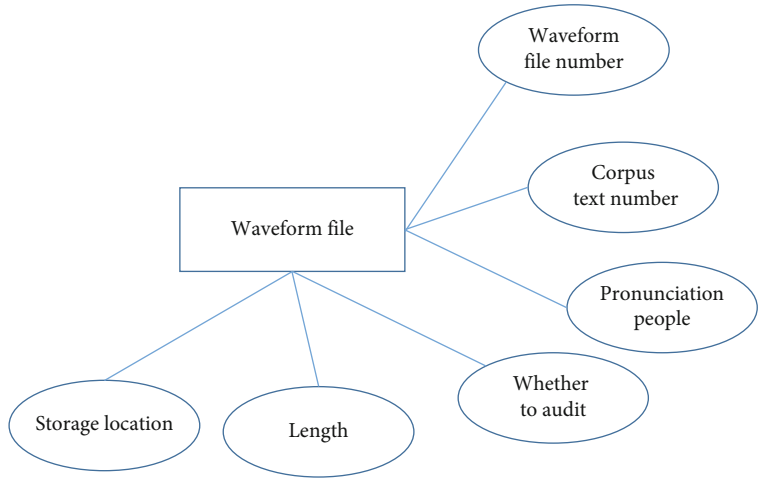

Figure 9: Pronunciation waveform file entity diagram.

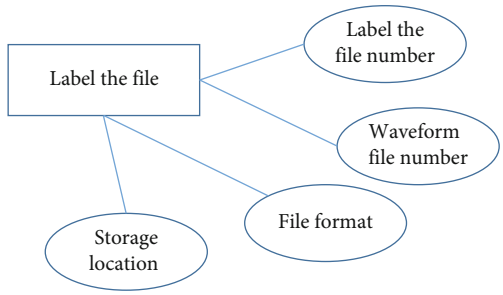

Figure 10: Labeling the file entity diagram.

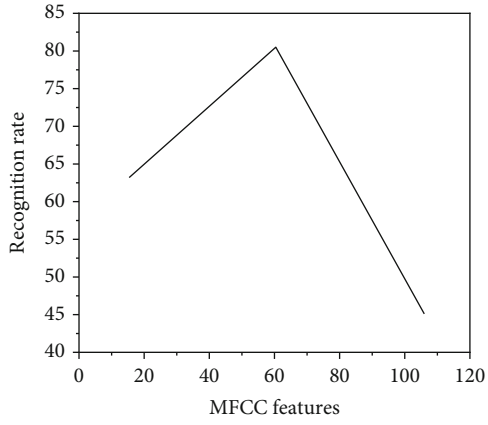

Figure 11: Comparison of recognition rates between MFCC features and DAE midlayer features (broken lines).

required. If the registration fails, the system prompts you to review the registration information and returns to the registration page. Its functional flow chart is shown in Figure 3.

(2) User Administrator Login Module. The login module ensures that users can log in to the system with legitimate identities and obtain the recording information for easy query and modification. First, the user fills in the account number and password as parameters and passes them to the server, which is compared with the information in the user information table [23]. If the authentication fails, an error message is displayed and the registration page is displayed. If the authentication succeeds, the user information and permission are displayed. Its function flow chart is shown in Figure 4.

(3) Recording Module. Recording module includes a recording program. After the program receives the user's request, according to the user's choice, the corresponding recording text is selected. The second step is to initialize the recording device on the machine and start recording. After the recording is complete, the user information, recording files, and text information are sent back to the server as parameters, and the save program is invoked for further processing.

(4) Database Maintenance Module. Database maintenance module is used to operate the database for administrators. The client provides an exchange interface, which is convenient for administrators to log in so as to manage and audit the user, corpus text, waveform files, and annotated files. The first step is to identity verification to see whether they have the authority to manage the database. After passing the verification, they can operate and manage the database and save the process of modifying information.

(5) Query and Download Module. The user sends a query request to confirm the user permission. Then, the query information input by users (articulator attributes, corpus text keywords, waveform file numbers, etc.) is transferred to the server as parameters. And the server returns the query result, and the user can select the corresponding file according to the returned result for download. Its function flow chart is shown in Figure 5.

3.2.2. The Conceptual Design of Database. The goal of the conceptual design is to accurately describe the information schema of the application domain and support the various applications of the user, so that it is easy to transform into database logic schema and easy to understand by the user. The typical method of conceptual model design is E-R

method. A diagram is composed of three parts, including entity, attribute, and connection. According to the requirement analysis of the system function and database described above, the E-R diagram of the system can be obtained, as shown in Figures 6–10.

3.2.3. System Implementation

(1) Implementation of Landing Module. In the process of system design, no matter if it is divided into several modules and different modules, its operators are different. The foundation of any successful application's security policy is a robust means of authentication and permission control and secure communications that provide data integrity and confidentiality. The design of login module is mainly to verify the correctness of user account and password.

(2) Implementation of Registration Module. It is mainly used for the personal information of registered users. Each user can view and modify their own information, as well as refer to previous personal recording records.

(3) Implementation of Recording Module. The first step is to initialize the recording device on the local device, and then, start recording after the user selects the corresponding text. After recording, the recording file is saved to the server.

(4) Implementation of Language Library Maintenance Module. It is mainly to achieve the management of the database, including the management of users, corpus text, waveform file management, and labeling file management.

(5) Implementation of Query and Download Module. Through the search function, users can find the corresponding text corpus, waveform files, and annotated files and download the required files.

Taking the tagging of speech corpus as an example, a complete speech corpus should not only contain original speech data and corresponding pronunciation text but also corresponding label files. In order to improve the utilization value of speech corpus, the key is to label the speech corpus completely. Corpus refers to a large-scale electronic text library scientifically sampled and processed. With the help of computer analysis tools, researchers can carry out relevant language theory and applied research. The label process of speech corpus is a process of language knowledge formalization. The label quality and depth of the speech corpus directly affect the accuracy and richness of information mined from the speech corpus and determine the availability and value of the speech corpus to a great extent. Based on statistical principles, we can find the habitual collocation of language, and the corpus can help us to better master the language. Is the tool for our research and collocation. A complete label system is a very important part of corpus construction, and the complete label includes segmenting and prosodic label. The so-called English phonetic segment annotation is to segment each phonetic unit (sentence, word, character, syllable, consonant, and vowels) in a continuous speech stream and describe their timbre characteristics, mainly including vowels, consonants, and combinations of vowels and vowels, vowels and consonants, and consonants and consonants.

A GMM-HMM acoustic model is simultaneously trained on HTK platform for the two features (unsupervised and supervised) and MFCC features obtained through deep autoencoder model training. The recognition accuracy of words and sentences is used as the experimental comparison results, as shown in Figure 11.

4. Conclusions

As one of the hottest research fields at present and in the future, deep learning has achieved good results in the field of speech recognition. The performance of speech recognition system often directly affects the effect experience of most intelligent systems, so the future development direction must be to combine the two technologies to promote mutual progress. Based on the theory of deep learning, the research comprehensively discusses the application value and effect of deep learning model in the field of speech recognition, starting from speech feature extraction and acoustic modeling.

(1) Taking acoustic feature extraction as the research object, the research work was carried out based on the deep autoencoder model. Deep autoencoders belonged to multilayer network model and were widely used in data dimension reduction and feature extraction based on unsupervised training. The research focused on the analysis of the deep learning model from the perspectives of feature data preprocessing, model structure, and network training parameters. The automatic encoder was established based on the speech features in MATLAB platform to extract the new speech features from the original MFCC features. Finally, HTK recognition tool was used to test and verify the TIMIT English speech corpus. Compared with the original situation, the new features extracted from the unsupervised and supervised training improved the English word recognition rate by 1.64% and 2.86% and the English sentence recognition rate by 2.55% and 6.53%, respectively

(2) Taking acoustic modeling as the research object, the research work was carried out based on DNN-HMM. As a discriminative model, deep neural network was applied in the field of acoustic modeling. It relied on the output layer to represent the HMM state output probability. And with the help of its own network structure, it could meet the requirements of complex feature modeling. It replaced the original GMM model and combined with HMM to obtain the acoustic model based on DNN-HMM. The acoustic models based on the GMM-HMM and DNN-HMM were modeled by Kaldi speech recognition system platform. Finally, experiments on TIMIT speech corpus prove that compared with

the GMM-HMM model, the English word recognition error rate and sentence recognition error rate of DNN-HMM model were reduced by 30.3% and 17.2%, respectively

Currently, with the rapid development of computer technology, English vocabulary and speech corpus recognition technology have also obtained the rapid development. And there are more and more technologies applied to the actual products, such as speech input system and computer assisted language learning system. Products are constantly emerging, which provides superior service for the people. For an excellent speech synthesis and recognition system, a speech corpus with high information content and low redundancy is essential. It can be seen that speech corpus plays an important role in speech recognition, speech synthesis, and other areas of speech research.

In the research, an English vocabulary and speech corpus was proposed and built to expand the sources of speech corpus and improve the efficiency of English vocabulary and speech corpus recognition and synthesis system construction. The following work were mainly completed.

(1) For English vocabulary and speech corpus selection, the original corpus text was automatically downloaded from the Internet firstly, and then, the greedy algorithm was used to screen the original corpus (based on high frequency words and three-tone words), and the final recorded corpus text was obtained

(2) For speech recording and corpus management system, the recording module working process and the design idea were described in detail. And the speech database management system was implemented, which was convenient for user to operate text, audio files, and tagging corpus query and download files. The recording work was tested, and the English vocabulary and speech corpus was established

(3) For speech file label, the speech corpus label standards and guidelines for corpus label in the United States were introduced. And then, preliminary speech label files were automatically generated by the program without alignment, which could effectively reduce the workload. And then through the software, the manual alignment work was performed and the final label files were obtained

Data Availability

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

Conflicts of Interest

The author declares that there are no conflicts of interest.

References

[1] V. H. Vu, Q. P. Nguyen, K. H. Nguyen, J. C. Shin, and C. Y. Ock, "Korean-Vietnamese neural machine translation with named entity recognition and part-of-speech tags," *IEICE Transactions on Information and Systems*, vol. 103, pp. 866–873, 2020.

[2] D. Lemmenmeier-Batinić, "Converting raw transcripts into an annotated and turn-aligned TEI-XML corpus: the example of the corpus of Serbian forms of address," *Slovenščina 2 0 Empirical Applied and Interdisciplinary Research*, vol. 9, no. 1, pp. 123–144, 2021.

[3] K. Zvarevashe and O. O. Olugbara, "Recognition of speech emotion using custom 2d-convolution neural network deep learning algorithm," *Intelligent Data Analysis*, vol. 24, no. 5, pp. 1065–1086, 2020.

[4] L. R. Kishline, S. W. Colburn, and P. W. Robinson, "A multimedia speech corpus for audio visual research in virtual reality (l)," *The Journal of the Acoustical Society of America*, vol. 148, no. 2, pp. 492–495, 2020.

[5] Y. Ahn, S. J. Lee, and J. W. Shin, "Cross-corpus speech emotion recognition based on few-shot learning and domain adaptation," *IEEE Signal Processing Letters*, vol. 28, pp. 1190–1194, 2021.

[6] J. Liu, W. Zheng, Y. Zong, L. U. Cheng, and C. Tang, "Cross-corpus speech emotion recognition based on deep domain-adaptive convolutional neural network," *IEICE Transactions on Information and Systems*, vol. E103.D, no. 2, pp. 459–463, 2020.

[7] W. Zhang, P. Song, D. Chen, C. Sheng, and W. Zhang, "Cross-corpus speech emotion recognition based on joint transfer subspace learning and regression," *IEEE Transactions on Cognitive and Developmental Systems*, vol. 14, pp. 588–598, 2021.

[8] W. Zheng, W. Zheng, and Y. Zong, "Multi-scale discrepancy adversarial network for crosscorpus speech emotion recognition," *Virtual Reality & Intelligent Hardware*, vol. 3, no. 1, pp. 65–75, 2021.

[9] L. Li and L. Cao, "Semantic analysis of literary vocabulary based on microsystem and computer aided deep research," *Mobile Information Systems*, vol. 2021, Article ID 8624147, 13 pages, 2021.

[10] S. P. Yadav, S. Zaidi, A. Mishra, and V. Yadav, "Survey on machine learning in speech emotion recognition and vision systems using a recurrent neural network (RNN)," *Archives of Computational Methods in Engineering*, vol. 29, no. 3, pp. 1753–1770, 2022.

[11] K. Chouhan, A. Shrivastava, C. Gangadhar, V. Shukla, and S. K. Jain, "Speech recognition classification with ANN implementation using machine learning algorithm," *Linguistica Antverpiensia*, vol. 2021, no. 1, pp. 2785–2796, 2021.

[12] M. Rojc and I. Mlakar, "An LSTM-based model for the compression of acoustic inventories for corpus- based text-to-speech synthesis systems," *Computers and Electrical Engineering*, vol. 100, article 107942, 2022.

[13] X. Ren, "Research on a software architecture of speech recognition and detection based on interactive reconstruction model," *International Journal of Speech Technology*, vol. 24, no. 1, pp. 87–95, 2021.

[14] O. Ivanova, J. J. Meilán, F. Martínez-Sánchez, I. Martínez-Nicolás, T. E. Llorente, and N. C. González, "Discriminating speech traits of Alzheimer's disease assessed through a corpus

of reading task for Spanish language," *Computer Speech & Language*, vol. 73, article 101341, 2022.

[15] I. Lefter, A. Baird, L. Stappen, and B. W. Schuller, "A cross-corpus speech-based analysis of escalating negative interactions," *Frontiers in Computer Science*, vol. 4, article 749804, 2022.

[16] S. Kibria, A. M. Samin, M. H. Kobir, M. S. Rahman, M. R. Selim, and M. Z. Iqbal, "Bangladeshi Bangla speech corpus for automatic speech recognition research," *Speech Communication*, vol. 136, pp. 84–97, 2022.

[17] A. Pandey and D. L. Wang, "Self-attending RNN for speech enhancement to improve cross-corpus generalization," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 1374–1385, 2022.

[18] L. Xia, G. Chen, X. Xu, J. Cui, and Y. Gao, "Audiovisual speech recognition: a review and forecast," *International Journal of Advanced Robotic Systems*, vol. 17, no. 6, 2020.

[19] C. M. Chen, M. C. Li, and M. F. Lin, "The effects of video-annotated learning and reviewing system with vocabulary learning mechanism on English listening comprehension and technology acceptance," *Computer Assisted Language Learning*, vol. 35, pp. 1557–1593, 2020.

[20] R. Baumann, K. M. Malik, A. Javed, A. Ball, B. Kujawa, and H. Malik, "Voice spoofing detection corpus for single and multi-order audio replays," *Computer Speech & Language*, vol. 65, article 101132, 2021.

[21] J. Basu, S. Khan, R. Roy, T. K. Basu, and S. Majumder, "Multilingual speech corpus in low-resource eastern and northeastern Indian languages for speaker and language identification," *Signal Processing*, vol. 40, no. 10, pp. 4986–5013, 2021.

[22] J. Gideon, M. G. McInnis, and E. M. Provost, "Improving cross-corpus speech emotion recognition with adversarial discriminative domain generalization (ADDoG)," *IEEE Transactions on Affective Computing*, vol. 12, no. 4, pp. 1055–1068, 2021.

[23] A. Vempala and E. Blanco, "Extracting biographical spatial timelines: corpus and experiments," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 1395–1403, 2020.