Hindawi
Mobile Information Systems
Volume 2022, Article ID 7186375, 12 pages
https://doi.org/10.1155/2022/7186375

Research Article

English Speech Recognition and Evaluation of Pronunciation Quality Using Deep Learning

Yushu Xu

Foreign Language Department, Zhejiang College of Shanghai Finance and Economics, Jinhua, Zhejiang 321013, China

Correspondence should be addressed to Yushu Xu; z2011219@shufe-zj.edu.cn

Received 19 February 2022; Revised 20 March 2022; Accepted 28 March 2022; Published 13 April 2022

Academic Editor: Abid Yahya

English is now one of the most important languages for economic exchange in various countries around the world, and it is also the most widely used language for cultural and information exchange. Like other countries, China likewise attaches highest significance to English learning, and people's demand for applied learning is also increasing rapidly these days. However, there are significant differences between Chinese pronunciation and English pronunciation, and China lacks an English environment while teaching English language. Furthermore, the traditional education is limited by the place and time of classes, due to which it cannot meet people's needs for learning English. With the fast progress of computer knowledge, the emergence of deep learning technology can better identify English pronunciation and evaluate the quality of English pronunciation. Additionally, deep learning can provide learners with precise, objective, and rapid pronunciation information. It can also assist learners in determining the differences between their pronunciation and conventional pronunciation through frequent listening and comparison, as well as correcting their pronunciation faults and increasing language learning efficacy. This study looks into the difficulty of using deep learning to evaluate the quality of English speech recognition and pronunciation. To evaluate English pronunciation quality, this paper selects intonation, speed, and rhythm, as the distinguishing indicators. The comparison between the results of manual evaluation and our evaluation clearly shows that English speech recognition and pronunciation quality model using deep learning established in this paper has much higher reliability. Among the 240 samples tested, only 32 samples differ by one grade, and the rest are similar.

1. Introduction

Trade throughout boundaries is growing fast in the age of economic globalization. English has attained extraordinary global importance as an international language. These days, with the rapid advancement of technology, communication between people and interaction between humans and devices are inextricably linked to information exchange. It has also become an essential component of human society. The rapid advancement of modern computer processing techniques encourages in-depth learning of speech recognition and enhances the accuracy of English voice quality assessment. Because people speak English with varying pronunciations, speeds, and intonation, this study employs a deep learning method to assess the recognition of English speech and pronunciation reliability. Many countries, which include China, encourage English learning. The claim for English learning [1] in China is increasing as a result of globalization and China's growing degree of internationalization. Due to period and place restrictions, there is an absence of a local English learning environment due to the great Chinese pronunciation features and the difference with English pronunciation. Furthermore, for a variety of reasons, good English teachers and standard classroom instruction [2] are unable to meet the English learning needs. Because of these factors, full English teaching and learning has become a major issue for people. English as a second language has grown in popularity as a research topic in the field of education. However, there are numerous flaws in English teaching as a second language for a variety of reasons. Many English

learners excel at listening, reading, and writing but struggle with speaking. They were able to overcome the difficulty thanks to computer-aided language learning.

Computer-aided language learning is the key to speech evaluation and recognition. Speech pronunciation varies greatly and has a large number of speech signals. The dimensions of the speech feature parameters are large, and the calculation of speech recognition and evaluation is large. It causes traditional speech recognition algorithms to become clogged. It is critical to use machine learning methods and collected big data for the improvement of speech recognition and quality. Speech recognition and evaluation technology have innovative dramatically in recent years, thanks to improvements in deep learning, cloud computing, and big data technologies. Deep learning (DL) is a branch of machine learning that tries to develop and imitate the human brain's deep neural network (DNN) for analysis and learning [3]. Since it can activate the neurons in the human brain to carry out multi-layer depth transmission to interpret data, DNN has demonstrated significant advantages in addressing various complex issues, which have been confirmed in the field of speech recognition. The computing complexity of DNN is no longer a problem, thanks to the rapid development of graphics calculators and cloud computing technology. As a result, research into deep learning-based English speech recognition technology can significantly increase the ability of speech information processing. Furthermore, it improves the efficacy of information acquisition and provides an improved user involvement [4].

Speech recognition technology research started in the 1950s. Bell Labs created ten isolated numerical identification systems in 1952 [5]. The authors of [6] planned the Hidden Markov Models (HMM) in the 1980s. The numerical model founded on this technique [7] regularly dominated speech recognition research. The HMM model accurately describes the short-term immobile features of speech signals and combines linguistics, acoustics, and syntax knowledge into a combined framework. HMM, investigation, and application have gradually gained traction [8]. The SPHINX system developed by the authors of [9] is the first "non-specific continuous speech recognition scheme." The GMM–HMM framework serves as the foundation, with the Gaussian Mixture Model (GMM) utilized to detect the chance of speech. HMM modeling simulates the effectiveness of speech. The ANN, DNN's predecessor, became a director of speech recognition research in the late 1980s [10]. This type of shallow neural network, on the other hand, has a broad effect on speech recognition functions and performs poorly compared to the GMM-HMM model. Keeping in view of the above, this research work utilizes the deep learning algorithm to establish the English speech recognition and pronunciation quality evaluation model using deep learning and takes the pitch, rhythm, speed, and intonation as the evaluation model indicators.

The following are the contributions of this paper to the research of English speech recognition and quality evaluation using deep learning:

(1) This study looks into the difficulty of using deep learning to evaluate the quality of English speech recognition and pronunciation. It selects intonation, speed, rhythm, and intonation as the indicators to evaluate English pronunciation quality

(2) Furthermore, using a deep learning method to extract high-level speech features, training a deep neural network, selecting a fixed hidden layer output as a new speech feature for a newly created network, and train GMM with novel speech characteristic

(3) Subsequently, a multi-parameter diagnostic model was established among 24 college students for the recognition of English speech and the quality of English pronunciation and was verified through simulation experiments

(4) Finally, it compared the results of the manual assessment and machine assessment, which proved that the proposed model has higher reliability and accuracy of the English speech recognition as well as pronunciation quality model using deeper learning than existing models

Left over of this paper is organized as follows: Section 2, illustrates scholars' works, related to the theme of the selected title, Section 3, explains signal preprocessing of speech and extraction of feature, Section 4, describes my concept and model based on deep learning, Section 5, clarifies the evaluation of English speech recognition and pronunciation quality using deep learning and lastly, Section 6 concludes my research work.

2. Related Work

The recognition of the English language is crucial in evaluating college English multimedia teaching [11, 12]. It is the technique for automatically translating speech signals into corresponding objects or words using machine learning and artificial intelligence [13]. Due to improvements in deep learning, big data, and cloud computing, speech recognition and evaluation criteria have improved quickly in recent years. One of the most widely spoken languages is English. The use of English as the main study object has transformed voice recognition technology into a hotspot of investigation. In the 1950s, AT&T Bell Labs Davis and other scholars developed a human oriented English digital recognition system based on the formant spectrum of digital vowels, namely, Audrey system [14]. The introduction of this system indicates the full application of automatic speech recognition system. At the same time, Lincoln Laboratory of MIT and RCA Laboratory of Princeton also focus on the analysis of speech recognition technology [15]. During this period, scholars studied the recognition system on the basis of small vocabulary, some groups of people and individual words. With the fast growth of computer technology, the basic hardware required for the technology of speech recognition is provided, and the emergence of linear-predictive-coding technology (LPC) and dynamic time planning (DTW) technology better deal with the problems existing in extracting accurate speech features, matching features, and templates

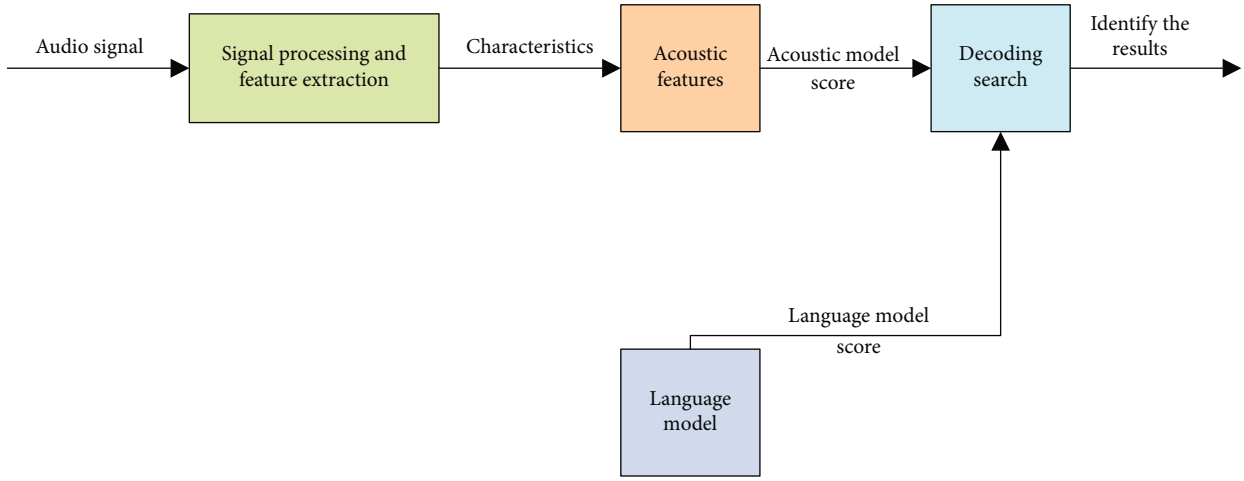

Figure 1: System structure of recognition of speech.

in speech recognition, and make full use of the recognition system in the market [16].

In the 1980s, speech recognition technology has achieved a lot of research results in language model and acoustic model. Sphinx is the world's first speaker-independent non-stop speech recognition system with a big vocabulary. The emergence of this system has realized the full application of HMM technology in acoustic model and achieved remarkable results [17]. The extraction of speech parameters, design model, and adaptive speech recognition technology was developed in the 1990s. At the same time, the technology is becoming more and more mature, and the technology is initially market-oriented. With the development of decades, traditional speech recognition technology is becoming more and more mature, and a complete recognition framework is constructed, including acoustic model based on GMM-HMM, speech feature extraction based on cepstrum and n-ary language model based on statistics [18–20]. In the twenty-first century, especially since 2009, the development of technology regarding recognition of speech has accelerated which is directly related to the rapid development of recognition of speech technology using deep learning, and accumulates a huge quantity of original speech data at the same time. Lee and other scholars take the modified linear function as the excitation function of hidden layer nodes, and use recursive neural network [21] in music processing. Li Deng et al. Recognized different languages and phonemes on the basis of multi-layer result conditional random fields [22]. The most famous achievement is the application of deep neural network in HMM state output probability modeling. Based on the fusion HMM model, a system for the recognition of speech using DNN-HMM acoustic model is constructed. Related with the conventional recognition of speech method with GMM-HMM as acoustic model, the results show that the word error rate (WER) of traditional speech recognition is about 30% higher [23]. Inspired from the work of above, this paper tries to design a model for English speech recognition and pronunciation quality evaluation founded on deep learning.

3. Signal Preprocessing of Speech and Extraction of Feature

3.1. Speech Recognition Technology. Once our vocal cords vibrate the air around them, they produce a series of sound waves, which we call speech. A microphone records sound waves, which are then converted into electrical signals. The signal is then separated into letters and words using advanced signal processing technology. Cheers to marvelous current progresses in artificial intelligence and machine learning, the machine can acquire to understand speech from involvement over time. But it is signal processing that enables all of this. In contrast, speech recognition is the capability of a machine or program to recognize words spoken aloud and alter them into readable form. Speech recognition employs a diverse set of research techniques from computer science, linguistics, and computer engineering. Many modern devices and text-focused programs include speech recognition functions, allowing for hands-free or simple device use. Speech recognition technology, on the other hand, allows computers to capture spoken sounds, understand them, and generate text from them. Users can easily control systems and create documents by talking thanks to technological progresses. Speech recognition allows for faster document creation because the software generates words as soon as they are uttered, which is comparatively quicker than an individual can type. Keyword recognition technology and continuous speech recognition technology are two main types of speech recognition technology. Figure 1 shows the most representative structure of a continuous speech recognition system. Its components include processing of signal and extraction of feature, acoustic model (AM), decoding search, and language model (LM).

3.2. Preprocessing of Speech Signal. Based division preprocessing is applied to the voice signal based on the pronunciation of language to reduce the amount of data supplied to the neural network's input, resulting in a significant increase in the sensitivity of the neural network's input data. Speech signal preprocessing is a critical stage in the construction

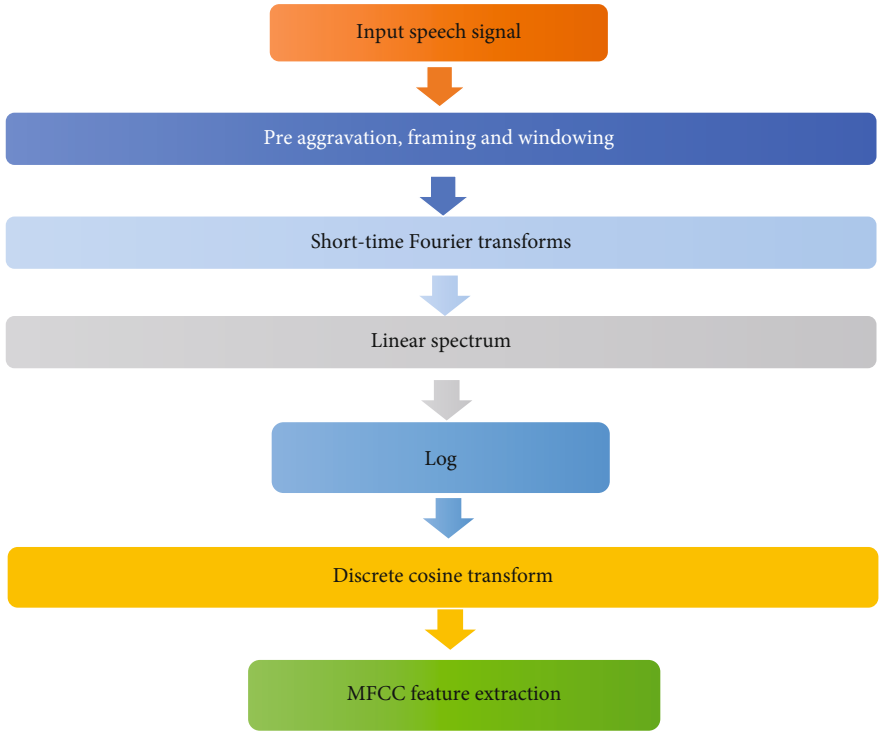

Figure 2: MFCC feature extraction flow diagram.

of an efficient and reliable speech recognition system. This stage consists of the following:

3.2.1. Pre-Aggravation. For speech signals, the high-frequency components will be reduced after the completion of short-time Fourier transform. At this time, pre-emphasis processing is required. The common methods are the following first-order high pass digital filters [24]:

$$H(z) = 1 - \mu z^{-1}. \tag{1}$$

In equation (1), μ is the pre-weighting coefficient, which is taken in the range of 0.9-1.

3.2.2. Framing and Windowing. Owing to differences in the spectral properties of the phone, changes in parody, and random changes in the sound path, speech is an unstable signal. However, the speech signal is assumed to be static at short intervals and is thus analyzed on these short-term windows. The basic feature of the voice signals is short-time characteristic and the voice signal will be input in overlapping sections. The value of the voice signal frame is taken in the range of 10-30 ms. The overlapping part between frames is called frame shift. Assuming that $w(n)$ is the function of the window and $S(n)$ is the voice signal, equation (2) shows the voice signal after windowing processing.

$$s_w(n) = w(n) \cdot s(n). \tag{2}$$

When the window function is a rectangular window, increasing the order causes the greatest overshoot on the pass-band near the discontinuity point, which is a Gibbs phenomenon. The window function with a short side lobe can be used to avoid this issue. The window function can be calculated using equation (3).

$$w(n) = \begin{cases} 0.56 - 0.47 \cos\left(\dfrac{2\pi n}{N-1}\right), & 0 \le n \le N-1 \\ 0, & \text{other} \end{cases}. \tag{3}$$

3.3. Speech Signal Feature Extraction. The classification problem lies at the heart of voice recognition, and the feature plays a significant role in it. The benefits and cons of the speech feature can be stated to influence the speech recognition system's performance improvement to some extent. The principal component analysis is used to obtain i-vectors. The session and speaker variability of the super vector has been extended to Joint Factor Analysis (JFA) [25]. The extracted i-vector captures the speaker and channel variations concurrently and effectively. The cosine kernel-based predicated process is used between the test speaker I vector and the target speaker I vector to detect any utterance in a target list. This method yields better optimum outcomes. Windowing the indication, putting on the DFT, taking the log of the magnitude, and then warping the frequencies on a Mel scale, followed by using the reverse DCT are all parts of the MFCC feature withdrawal technique. At present, most speech recognition technologies need to extract speech signal features before application. The extracted features mainly include I-vector and Mel-frequency cepstral coefficients (MFCC) [26]. Figure 2 shows the overall process of the MFCC feature extraction.

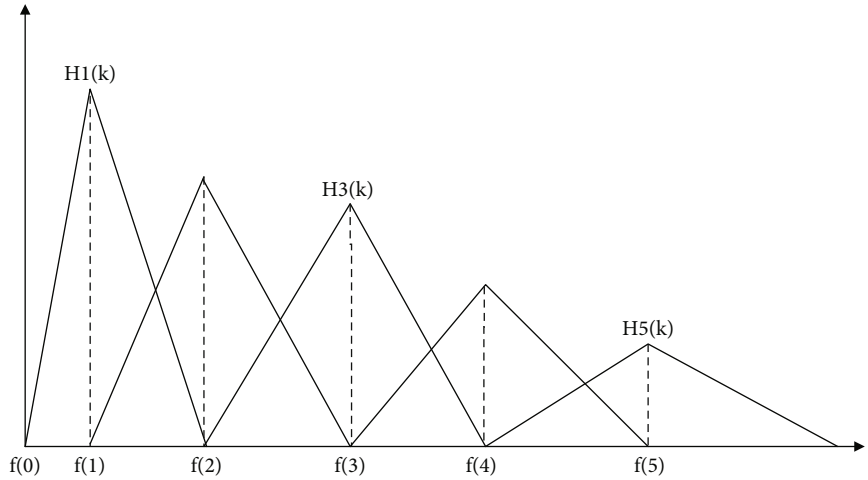

Figure 3: Mel filter bank.

The following is the MFCC feature extraction process:

(i) By preprocessing the input speech signal, the time-domain signal with frame and window can be obtained

(ii) The corresponding linear spectrum can be obtained by short-time Fourier transform of all windowed time-domain speech signals, as by equation (4).

$$S_i(k) = \sum_{n=0}^{N-1} s_i(n) e^{-j\frac{2\pi nk}{N}} (0 < K < N). \quad (4)$$

MFCC features are constructed on the basis of human auditory perception, so the linear frequency should be converted to the corresponding human ear frequency. Equation (5) is the association among general and Mel frequencies.

$$\text{mel}(f) = 2595 \lg\left(1 + \frac{f}{700}\right). \quad (5)$$

(iii) The energy spectrum passes through a group of triangular bandpass filters with 2429 filters. $F(m)$ represents the frequency of the bandpass filter, the value of M is 1, 2, ... $F(m)$, and the value of m gradually decreases, and the interval distance continues to shorten, which can be represented by Figure 3. The triangular bandpass filter is defined by equation (6).

$$Hm(k) = \begin{cases} \dfrac{k - f(m-1)}{f(m) - f(m-1)}, & f(m-1) \le k \le f(m) \\ \dfrac{f(m+1) - k}{f(m+1) - f(m)}, & f(m) \le k \le f(m+1) \\ 0, & \text{other} \end{cases}, \quad (6)$$

$$\sum H_m(k) = 1. \quad (7)$$

In the above equation, m is the filter number, while $f(m-1)$ represents the frequency's lower limit of the filter, $f(m)$ is the center frequency of the filter, and $f(M+1)$ represents the frequency's upper-limit of the filter. The filter can make the spectrum smoother, avoid the interference caused by harmonics, and reduce the amount of characteristic data.

(iv) By utilizing equation (8), we can calculate the logarithmic energy E_m of every group of filters

$$E_m = \text{In}\left(\sum_{k=0}^{N-1} P(k) H_m(k)\right). \quad (8)$$

3.4. Acoustic Model. The relationship between an audio signal and the phonemes or other linguistic units that make up speech is represented by an acoustic model in automatic speech recognition. A set of audio recordings and their transcripts are used to train the model.

3.4.1. Acoustic Model Modeling Unit. The corresponding monitoring unit, which includes vowels, phonemes, syllables, half syllables, and words, should be picked first when creating the acoustic model unit. Usually, in the process of Chinese recognition, conference vowels are used as modeling units. Single-factor modeling can be used during modeling, and elements altering context pronunciation can be considered to achieve three-factor modeling. After successfully establishing the three-factor model, bind the state and reduce the number of training acoustic units. Figure 4 shows the probability parameters of hidden Markov.

3.4.2. GMM-HMM Acoustic Modeling. Fresh machine learning methods in automatic voice recognition can lead to significant advancements. According to [26], the most significant single development was the advent of the Expectation-Maximization technique for training hidden Markov models (HMMs) approximately four decades ago.

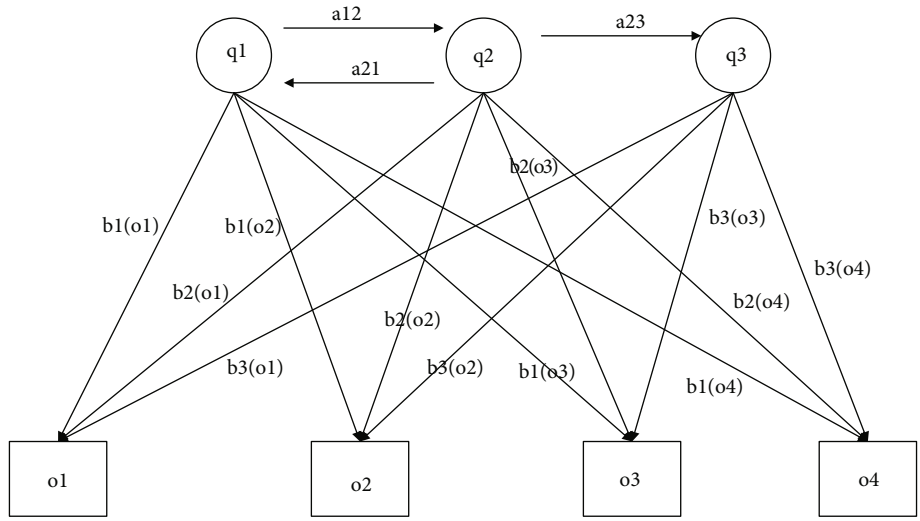

Figure 4: Probability parameters of hidden Markov.

The EM technique made it possible to use the richness of Gaussian mixture models (GMM) to construct speech recognition systems for real-world tasks [27]. This makes it easier to visualize the connection between HMM states and acoustic input. GMMs offer several properties that make them ideal for modeling probability distributions over vectors of input attributes associated with each HMM state. They can represent posterior distribution to any desired level of accuracy with enough components, and they are relatively easy to fit data using the EM technique. Furthermore, the hidden Markov model can create a sequence model based on speech acoustic features, making it useful for speech recognition. A state transition probability matrix, an a priori probability vector, and a Gaussian mixture model of relevant states are combined to generate the GMM-HMM parameter set. The field on phonemes is related to the state on GMM-HMM. The hidden Markov model is depicted in Figure 3. The output probability distribution model is considered to be GMM based on the hidden Markov probability parameters. The homogeneous Markov chain's transition probability matrix is shown in Equation (9).

$$aij = P(q_t = j | q_{t-1} = i), i, j = 1, 2, \cdots, N. \tag{9}$$

The initialization probability of the Markov chain is represented by equation (10)).

$$\pi = [\pi_i], i = 1, 2, \cdots, N. \tag{10}$$

In problem-solving process of speech processing, the probability distribution of continuous observation is explained by HMM probability density function, in which the Gaussian mixture function is the probability density function with the widest application range as by equation (11).

$$b_i(o_t) = \sum_{m=1}^{M} \frac{c_{i,m}}{(2\pi)^{D/2} |\sum_{im}|^{1/2}} \exp\left[-\frac{1}{2}(o_t - \mu_{i,m}) T \sum_{i,m}^{-1} (o_t - \mu_{i,m})\right]. \tag{11}$$

After the m-mixture component drops to 1, the output probability distribution based on this state degenerates into equation (12) Gaussian distribution.

$$b_i(o_t) = \frac{1}{(2\pi)^{D/2} |\sum_i|^{1/2}} \exp\left[-\frac{1}{2}(o_t - \mu_i)^T \sum_i^{-1} (o_t - \mu_i)\right]. \tag{12}$$

3.5. Language Model. Language model (LM) is a probabilistic statistical model, which can output hypothetical word sequences. The language model extensively used in the recognition of speech is n-gram. The essence of this model is a two-word language model, which fixes several historical words, predicts the probability of the current word, and scores the probability of outputting sentences [28].

Assuming that h represents the sequence of historical words, calculate the probability $P(w|h)$ of the occurrence of W words. The common way to estimate the probability of current words is in statistical word frequency. If the number of sentences is large, the corresponding occurrence of statistical historical word sequence c_1 is more, and c_2 is the sequence of historical words and the number of current words, which is expressed by equation (13).

$$\mathrm{P}(w|h) = \frac{c_1}{c_2}. \tag{13}$$

When the number of sentences is small, it is difficult to use the current sentence quantity statistical word frequency method. In case of c, it is difficult to calculate the probability of $P(w|h)$. Therefore, the length of w_1^n word sequence is n and the corresponding joint probability is $P(w_1, w_2, \cdots, w_n)$. Based on the chain rule, the joint probability is obtained by equation (14).

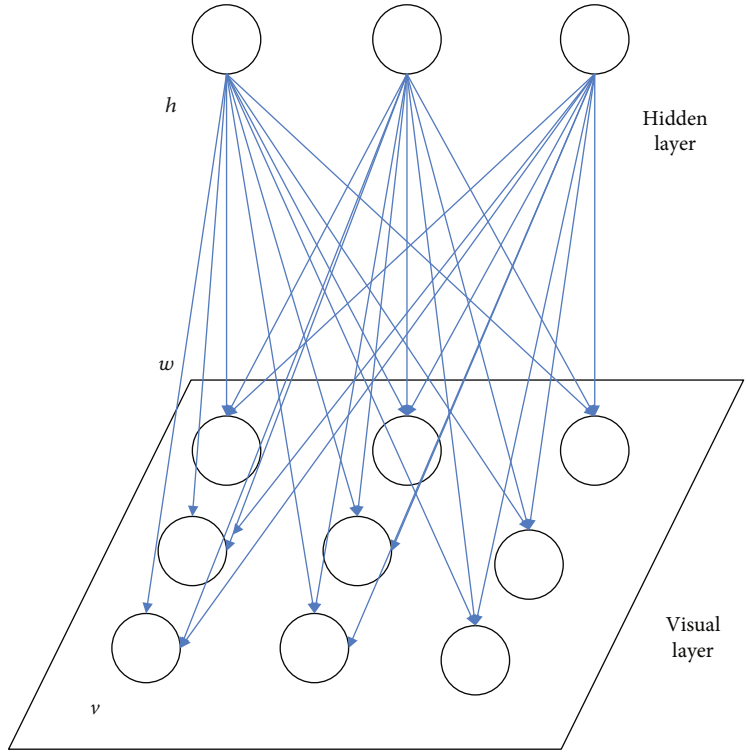

Figure 5: RBMs model network topology.

$$P(w_1^n) = P(w_1)P(w_2|w_1)P(w_3|w_1^2)\cdots P(w_n|w_1^{n-1}) = \prod_{k=1}^{n} P\left(w_k\middle|w_1^{k-1}\right). \tag{14}$$

4. Concept and Model Based on Deep Learning

4.1. Deep Learning Concept. Deep learning is a type of machine learning method that is based on artificial learning neural networks. Deep learning is extremely powerful due to its ability to process a large number of features while dealing with unstructured data. Deep learning algorithms, on the other hand, can perform extremely well for less complicated issues because they need access to a huge quantity of data to be effective. Based on the traditional neural network model, a deep learning model is formed, in which there are many artificial neural networks with different hidden layers. Deep learning can realize artificial intelligence tasks with strong abstraction. The common types of tasks are image recognition, speech recognition, image retrieval, and natural language understanding. The deep learning structure has a direct impact on its ability to model and express features. Furthermore, when compared to traditional networks, deep learning's expression and modeling abilities are superior. Neural networks can be distributed into 2 categories: unsupervised learning and supervised learning. Deep learning is the same, but the difference lies in the different learning frameworks and the corresponding learning model.

4.2. Deep Learning Model. Deep learning is the process by which a computer design acquires to conduct organization tasks straight from sound, images, or text. Deep learning models can attain high accuracy, sometimes outperforming humans. This section explains restricted Boltzmann machines (RBMs) and deep belief networks (DBNs) of my proposed model.

4.2.1. RBMs Model. Restricted Boltzmann machines (RBMs) belong to stochastic neural networks, which generate corresponding network neuron states based on different probability methods. From this, it can be concluded that the neuron states obey [29] to the state distribution samples. At present, the back-propagation network (BP) is widely used in many kinds of artificial neural network models, which directly affects the artificial neural network. As a multi-layer feed-forward model, the back-propagation model network is completed by using the error back-propagation algorithm during training [30]. The components of the BP neural network include the hidden layer, input layer, and output layer, which are shown in Figure 5. BP neural network uses activation function to explain the relationship between layers and simulates the interaction between different neurons based on activation function.

4.2.2. DBNs. Deep belief networks (DBNs) are also random deep neural networks. We can complete the statistical distribution, represent the abstract features of things, and establish the statistical model of things using this model [31, 32]. In the process of creating a recognition model of speech, this paper uses a network of deep belief to replace the previous Gaussian mixture model (GMM), which has a remarkable effect.

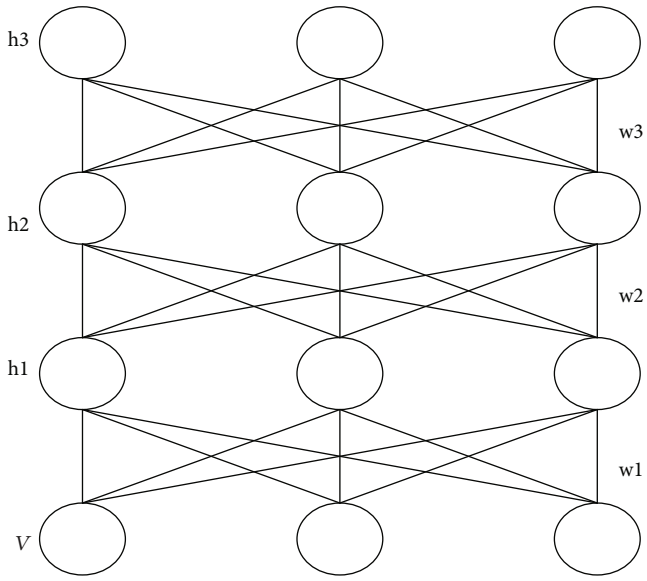

Figure 6: RBMs network.

The deep Boltzmann machine network can be gained by adding more hidden layers based on the RBMs network. In addition, the deep Boltzmann machine network is further improved. The model of a directed graph is adopted in the part close to the view, and the RBMs network can be utilized in the furthest visual layer [33]. The specific structure is shown in Figure 6.

By combining multiple RBMs layers to form DBNs and stacking multiple RBMs, neural networks with different depth levels can be established, and DBNs can be regarded as a generic model or as a discrimination model.

5. Evaluation of the English Speech Recognition and Pronunciation Quality Using Deep Learning

5.1. English Sentence Data Source. When measuring the quality of English speech recognition and pronunciation based on extensive learning, this paper selects 24 college students, and the numbers of girls and boys are 9 and 15, respectively. The tester uses CoolEdit recording software to complete the recording of English sentences with a frequency of 16 KHz. The recorded sentences are mainly several sentences commonly used in people's English expressions [34], as shown below:

(1) It will be in the same spot it's always been

(2) The clothing is stored in the refrigerator

(3) She intends to submit it on Wednesday

(4) I will be able to tell him tonight

(5) Along with the piece of wood, there's a black sheet of paper

(6) The team leader will arrive in seven hours

(7) What are the bags that are stacked beneath the table?

(8) They just carried it upstairs and are now bringing it back down

(9) I always travel home on weekends to see Agnes

(10) I simply want to get this out of the way and go out for a drink with Karl

5.2. English Speech Recognition and Pronunciation Quality Evaluation Index. When recognizing English speech and evaluating pronunciation quality, it is usually necessary to analyze the speaker's length, intonation, and sound play to complete the comprehensive evaluation. When reading paragraphs and sentences, we should judge the content of the tester's pronunciation and expression, and the phonological characteristics also have a decisive impact on the true meaning of the sentence. When evaluating the pronunciation quality of English sentences and paragraphs, it is necessary to judge whether the speaker can accurately master the core vocabulary of the sentence, whether it can distinguish weak reading or unimportant data, and whether the pronunciation length of the tester is accurate. That is, in the process of evaluating the pronunciation quality of English sentences, a high-quality pronunciation must first be able to accurately and completely read the sentences, and the pronunciation should be smooth and clear without wrong pronunciation. At the same time, the English speaking speed is appropriate and can accurately emit stress. To summarize, the evaluation indexes chosen for English speech recognition and pronunciation quality evaluation are pitch and rhythm.

5.3. English Speech Recognition and Pronunciation Quality Evaluation Model Based on Deep Learning

5.3.1. Speech Evaluation. The purpose of evaluating English speech recognition and pronunciation quality is to test the evaluation model and performance of English speech pronunciation quality established in this paper, that is, the same English sentence is evaluated by machine and man to judge whether the two sentences are consistent. Firstly, the reliability of manual evaluation is tested and then based on the reliability of the manual evaluation; finally, I compared the consistency between manual evaluation and machine evaluation. When expressing the consistency between manual evaluation and machine evaluation, the indexes selected in this paper are adjacent consistency rate, consistency rate, and Pearson's correlation coefficient. We can calculate the consistency rate by utilizing equation (15).

$$A_{\text{consistency rate}} = \text{Number of all samples that are consistent between machine evaluation and manual evaluation/Total number of samples.} \quad (15)$$

Table 1: Artificial evaluation grade and evaluation standard.

Level	Intonation	The speed	Rhythm	Tone of voice	As a whole
A	Accurate content, clear and fluent pronunciation, no mispronunciation	The speed is moderate	Precise accent, strong sense of rhythm	The tone is natural and precise	The overall pronunciation is excellent
B	The content is relatively complete, and the pronunciation is fluent and clear, without serious pronunciation mistakes	The speed faster	Accurate accent and good sense of rhythm	Intonation is relatively natural and accurate	Overall good pronunciation
C	The content is generally complete, and the pronunciation is basically fluent and clear, without interfering pronunciation	Speak too fast	The accent is common and has a sense of rhythm	The intonation is roughly accurate and unnatural	General pronunciation
D	The content is generally complete, and the pronunciation is basically fluent and clear, without interfering pronunciation	Speed super-fast	There is wrong accent pronunciation, less or more accent, no sense of rhythm	The intonation is unnatural and inaccurate	Poor overall pronunciation

The adjacent consistency rate is one grade lower than the difference between machine evaluation and manual evaluation. This can be calculated by utilizing equation (16).

$$A_{\text{adjacent consistency rate}} = \text{Number of samples consistent with manual evaluation and machine evaluation} + \text{Number of adjacent samples for machine evaluation and manual evaluation/Total number of samples.} \tag{16}$$

The statistics of linear correlation degree between two variables are expressed by Pearson's correlation coefficient, i.e., r, r represents the linear degree between different variables, and the value range is -1 to +1. If the absolute value is high, it shows that there is a solid correlation. Generally, $0 < r < 0.2$ is a very weak correlation, $0.2 < r < 0.4$ is a weak correlation, medium correlation is $0.4 < r < 0.6$, and the value range of solid correlation is $0.6 < r < 0.8$; the range of extremely strong correlation is $0.8 < R < 1$ [35].

5.3.2. Manual Evaluation. Starting from the characteristics of college students' English pronunciation quality and according to the suggestions given by English phonetics scholars, four grades are set for the evaluation results of sound speed, intonation, and rhythm, and the different grades are listed in Table 1.

This paper invites two teachers with rich experience in English teaching to evaluate the ten common English accents recorded by 24 college students in this university. The evaluation reference standards are speed, intonation, intonation, and rhythm. The final overall evaluation result is obtained according to the results of each index.

5.4. Deep Learning-Based Results for English Speech Recognition and Pronunciation Quality Evaluation. This manuscript adopts the deep learning algorithm to recognize English pronunciation and evaluate the English pronunciation quality of 24 college pupils participating in this test. Table 2 lists the evaluation index results of same sample. Here, I have taken 4 indicators such as intonation, speed, rhythm, and tone of voice. I further divided them into different levels and obtained their corresponding consistencies.

Table 3 lists the evaluation index results of same statistical index. Here, I have also taken 4 indicators such as intonation, speed, rhythm, and tone of voice. I obtained consistency rate, contiguous uniformity and Pearson of each indicator.

Table 2: Evaluation index results of same sample.

Indicators	Consistent	Level difference	Difference between the secondary	Level 3 difference
Intonation	208	33	1	0
Speed	196	42	2	0
Rhythm	205	35	4	0
Tone of voice	193	45	3	0

There are 208 levels in total for evaluation intonation when comparing manual and machine evaluation. The number of samples with 1 level change is only 33. There is one sample with 2 levels change, and there is no difference in the number of samples with three levels. The data shows that the consistency rate of pitch between manual evaluation and machine evaluation is 86.3%, the adjacent consistency rate is 99.59%, and the corresponding coefficient of Pearson's correlation is 0.9. The findings illustrate that the technique of pitch evaluation in this research work has high reliability. In the results of evaluating English speaking speed, 196 samples in the manual evaluation and machine evaluation are consistent, 42 samples are one level worse, and there are no two or three levels worse. The consistency rate, adjacent consistency rate, and Pearson's correlation coefficient of manual evaluation and machine evaluation are 82.1%, 100%, and 0.51, respectively. This suggests that the method of evaluating the English speech speed used in this paper is reliable.

When evaluating English speech rhythm, 205 sample grades in the manual evaluation and machine evaluation are consistent, 35 sample grades differ by one, three sample grades differ by two, and there is no sample with three grades. The manual and machine evaluation results show that the consistency rate, adjacent rate of consistency, and Pearson's correlation coefficient are 80.1 percent, 98.34 percent, and 0.63 percent, respectively. As a result, the method

Table 3: Evaluation index results of same statistical index.

Indicators	Consistent rate	Contiguous uniformity	Pearson
Intonation	86.3%	99.59%	0.9
Speed	82.1%	100%	0.51
Rhythm	84.9%	98.76%	0.55
Tone of voice	80.1%	98.34%	0.63

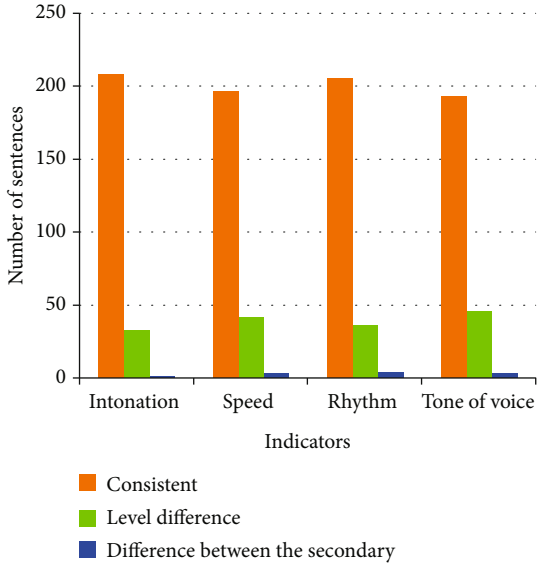

Figure 7: Overall evaluation results of machine and labor.

of evaluating English rhythm is reliable, and the results show that the method of evaluating intonation is reliable. Based on the analysis presented above, it is concluded that the evaluation of English speed, intonation, intonation, and rhythm used in this paper is feasible, and it is possible to develop an evaluation model for the quality of English pronunciation.

5.5. Test and Evaluation Model. As per the aforementioned evaluation indicators, this research work deeply analyzes the indicators and weights of English pronunciation, such as speed, intonation, intonation, and rhythm, and establishes the evaluation model of English pronunciation quality using regression analysis. The dependent variable is the total score of manual evaluation, and the independent variable is speed, intonation, intonation, and rhythm. Select and evaluate the same English sentences as manual evaluation. Based on multiple linear regression analyses and SPSS software, different evaluation index weights can be obtained, which are expressed by equation (17).

$$\begin{aligned} \text{Score} = {} & \text{AccuracyScore} \times 0.45 + \text{SpeedScore} \times 0.17 \\ & + \text{RhythmScore} \times 0.36 + \text{IntonationScore} \\ & \times 0.32\text{-}0.402. \end{aligned} \tag{17}$$

In addition, I select two different statistical tests to test the importance of the regression equation. The first is an

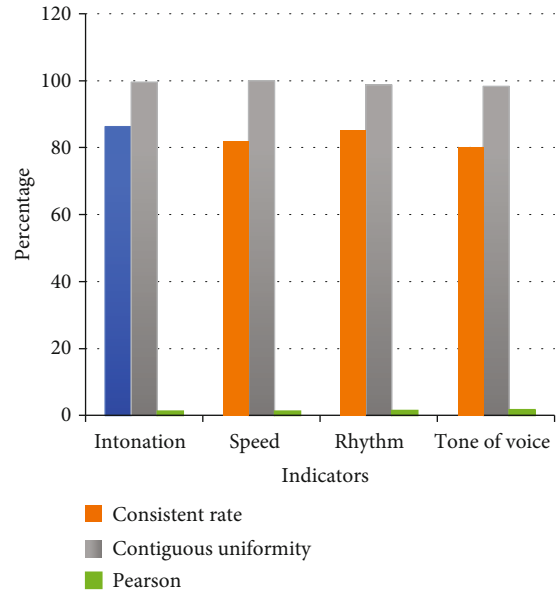

Figure 8: Comparisons of results of evaluation index of same statistical index.

F-test to determine the significance of the regression equation, and the second is a t-test to determine the significance of the regression coefficient. Figure 7 shows the results of using equation (17) to evaluate the overall situation of 24 students reading ten sentences. Among them, 208 samples have the same manual evaluation level and machine evaluation level, and 32 samples have a difference of one level, no difference of two poles, or three levels.

Figure 7 shows the results of the evaluation index for the same statistical index. According to this figure, the consistent rate of intonation, speed, rhythm, and tone of voice is 86.3%, 82.1%, 84.9%, and 80.1%, respectively. Similarly, their contiguous uniformities are 99.59%, 100%, 98.76, and 98.34%, respectively.

Figure 8 shows that the consistency rate of pitch between manual evaluation and machine evaluation is 86.3%, the adjacent consistency rate is 99.59%, and the corresponding coefficient of Pearson's correlation is 0.9. The findings illustrate that the technique of pitch evaluation in this research work has high reliability. In addition, in the results of evaluating English speaking speed, 196 samples in the manual evaluation and machine evaluation are consistent, 42 samples are one level worse, and there are no two or three levels worse. Pearson's correlation coefficient of consistency rate, adjacent consistency rate, and manual evaluation and machine evaluation are 82.1%, 100%, and 0.51, respectively, which indicates that the method of evaluating the English speech speed used in this paper is reliable.

6. Conclusions

These days, the rapid growth of big data, the technology of deep learning, and cloud computing have accelerated the development of speech recognition and evaluation. Deep learning simulates and learns the analysis process of the

human brain to form a deep neural network (DNN). Additionally, this technology can better transfer simulation and interpretation data to human brain neurons at multiple levels and improved the data processing speed. With globalization and China's increasing internationalization, the demand of Chinese people for English learning is increasing rapidly. This research work utilizes the deep learning algorithm to establish the English speech recognition and pronunciation quality evaluation model using deep learning and takes the pitch, rhythm, speed, and intonation as the evaluation model indicators. Comparing the results of voice evaluation and manual evaluation, it was found that out of 240 voice data, only 32 samples had grade differences, and the rest were identical. In addition, the results indicate that English speech recognition technology based on neural learning not only improves speech processing ability but also improves English pronunciation quality.

Data Availability

All the data is available in the paper for publication of this work.

Conflicts of Interest

I declare that there is no conflict of interest for publication of this paper.

References

[1] I. W. Suryasa, I. G. P. A. Prayoga, and I. W. A. Werdistira, "An analysis of students motivation toward English learning as second language among students in Pritchard English academy (PEACE)," *International Journal of Social Sciences and Humanities*, vol. 1, no. 2, pp. 43–50, 2017.

[2] X. Zhang and L. Chen, "College English smart classroom teaching model based on artificial intelligence technology in mobile information systems," *Mobile Information Systems*, vol. 2021, 2021.

[3] Z. Z. Chen and L. Yan, "Autonomous learning of College English under the network environment," *Journal of Southwest Agricultural University (Social Sciences Edition)*, vol. 10, pp. 128–131, 2011.

[4] H. Wu, "Multimedia interaction-based computer-aided translation technology in applied English teaching," *Mobile Information Systems*, vol. 2021, 2021.

[5] L. Schillingmann, J. Ernst, V. Keite, B. Wrede, A. S. Meyer, and E. Belke, "AlignTool: the automatic temporal alignment of spoken utterances in German, Dutch, and British English for psycholinguistic purposes," *Behavior Research Methods*, vol. 50, no. 2, pp. 466–489, 2018.

[6] A. Coutrot, J. H. Hsiao, and A. B. Chan, "Scanpath modeling and classification with hidden Markov models," *Behavior Research Methods*, vol. 50, no. 1, pp. 362–379, 2018.

[7] Z. Ali, A. W. Abbas, T. M. Thasleema, B. Uddin, T. Raaz, and S. A. R. Abid, "Database development and automatic speech recognition of isolated Pashto spoken digits using MFCC and K-NN," *International Journal of Speech Technology*, vol. 18, no. 2, pp. 271–275, 2015.

[8] E. Bocchieri, "System and method for speech recognition modeling for mobile voice search," *Jersey Citynj Usphiladelphiapa Uschathamnj Us*, vol. 47, no. 10, pp. 4888–4891, 2017.

[9] M. Telmem and Y. Ghanou, "Estimation of the optimal HMM parameters for amazigh speech recognition system using CMU-Sphinx," *Procedia Computer Science*, vol. 127, pp. 92–101, 2018.

[10] S. M. Siniscalchi and V. M. Salerno, "Adaptation to new microphones using artificial neural networks with trainable activation functions," *IEEE Trans Neural Netw Learn Syst*, vol. 28, no. 8, pp. 1959–1965, 2017.

[11] L. He, G. Jin, and S. B. Tsai, "Design and implementation of embedded real-time English speech recognition system based on big data analysis," *Mathematical Problems in Engineering*, vol. 2021, 2021.

[12] M. Sarma, P. Ghahremani, D. Povey, N. K. Goel, K. K. Sarma, and N. Dehak, "Emotion Identification from Raw Speech Signals Using DNNs," in *in Proceedings of the Interspeech*, pp. 3097–3101, Hyderabad, India, 2018.

[13] C. Cai, Y. Xu, D. Ke, and K. Su, "A fast learning method for multilayer perceptrons in automatic speech recognition systems," *Journal of Robotics*, vol. 2015, 7 pages, 2015.

[14] T. Xi, "Design of English diagnostic practice sentence repetition recognition system based on matching tree and edge computing," *Wireless Communications and Mobile Computing*, vol. 2021, 2021.

[15] J. G. Liu and Y. Zhou, "Estimation algorithm of switching speech power spectrum for automatic speech recognition system," *Journal of Computer Applications*, vol. 36, no. 12, 2016.

[16] Q. Y. Wang, R. Y. Liang, and L. Zhao, "Research on teaching and experimental methods of speech signal processing under the real-time environment," *Research and exploration in The Laboratory*, vol. 38, no. 9, 2019.

[17] G. Saon and J. T. Chien, *Large-Vocabulary Continuous Speech Recognition Systems*, vol. 54, no. 2, 2012Tsinghua University Press, 2012.

[18] W. Hu, Y. Qian, F. K. Soong, and Y. Wang, "Improved mispronunciation detection with deep neural network trained acoustic models and transfer learning based logistic regression classifiers," *Speech Communication*, vol. 67, pp. 154–166, 2015.

[19] D. Anggraeni, W. Sanjaya, and M. Munawwaroh, "Control of robot arm based on speech recognition using Mel-Frequency Cepstrum Coefficients (MFCC) and K-Nearest Neighbors (KNN) method," in *2017 International Conference on Advanced Mechatronics, Intelligent Manufacture, and Industrial Automation (ICAMIMIA)*, pp. 217–222, Surabaya, Indonesia, 2017.

[20] A. Dd, A. Aa, and B. Mg, "A new modeling approach for mixture fraction statistics based on dissipation elements," *Proceedings of the Combustion Institute*, vol. 38, no. 2, pp. 2681–2689, 2021.

[21] W. D. Lee, D. H. Kim, and S. G. Kang, "Noninformative priors for linear function of parameters in the lognormal distribution," *Journal of the Korean Data & Information Science Sociaty*, vol. 27, no. 4, pp. 1091–1100, 2016.

[22] P. Werther and S. Röder, *Method for carrying out a multimedia communication based on a network protocol, particularly TCP/IP and/or UDP.*, US, 2013.

[23] Z. G. Q. D. Fan, H. Li, and W. L. Zhang, "Hybrid language model speech recognition method based on MTL-DNN

system combination," *Journal of Data Acquisition & Processing*, vol. 32, no. 5, pp. 1012–1021, 2017.

[24] J. A. Cui and N. Xu, "Design and realization of a digital filter in a real-time audio signal acquisition system," *Acta scientiarum naturalium universitatis neimongol*, vol. 33, no. 2, pp. 277–280, 2010.

[25] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 19, pp. 788–798, 2001.

[26] J. Baker, L. Deng, J. Glass et al., "Developments and directions in speech recognition and understanding, part 1 [DSP Education]," *Signal Processing Magazine, IEEE*, vol. 26, no. 3, pp. 75–80, 2009.

[27] B. H. Juang, S. Levinson, and M. Sondhi, "Maximum likelihood estimation for multivariate mixture observations of Markov chains (Corresp.)," *IEEE Transactions on Information Theory*, vol. 32, no. 2, pp. 307–309, 1986.

[28] Z. R. Shi and J. X. Chen, "Event detection via recurrent and convolutional networks based on language model," *Journal of Xiamen University (Natural Science)*, vol. 58, no. 3, pp. 442–448, 2019.

[29] J. Yang Y. D. Sun et al., "Weakly supervised learning with denoising restricted Boltzmann machines for extracting features," *Acta Electronica Sinica*, vol. 12, pp. 2365–2370, 2014.

[30] S. Wang and X. Shi, "Research on correction method of spoken pronunciation accuracy of AI virtual English reading," *Advances in Multimedia*, vol. 2021, 12 pages, 2021.

[31] Y. Z. Zhao and W. B. Liu, "Research on unconstrained face recognition based on DBNs network," *Acta Metrologica Sinica*, vol. 38, no. 1, pp. 65–68, 2017.

[32] X. X. Wang, P. Chen, P. Liu, and M. L. Liu, "Geography ontology fusion model based on statistical machine learning," *Journal of tongji university (natural science)*, vol. 39, no. 5, pp. 758–763, 2011.

[33] S. S. Shao and L. B. Liu, "Improved deep belief network prediction model and its application," *Journal of Computer Applications*, vol. 38, no. z1, 2018.

[34] S. Zhang, *Research on the Video Production Methods of Micro-lessons*, A Study Based on Computer Technology, 2017.

[35] G. P. Chen and H. A. Wang, "Personalized recommendation algorithm on improving Pearson correlation coefficient," *Journal of Shandong Agricultural University (Natural Science Edition)*, vol. 47, no. 6, pp. 940–944, 2016.