Hindawi
Security and Communication Networks
Volume 2022, Article ID 8961836, 11 pages
https://doi.org/10.1155/2022/8961836

Research Article

Standardized Evaluation Method of Pronunciation Teaching Based on Deep Learning

Xiaoda Zhao and Xiaoyan Jin

Northeast Normal University, Changchun 130024, China

Correspondence should be addressed to Xiaoda Zhao; zhaoxd903@nenu.edu.cn

Received 7 January 2022; Revised 2 February 2022; Accepted 4 February 2022; Published 7 March 2022

Academic Editor: Muhammad Arif

Copyright © 2022 Xiaoda Zhao and Xiaoyan Jin. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

With the advancement of globalization, an increasing number of people are learning and using a common language as a tool for international communication. However, there are clear distinctions between the native language and target language, especially in pronunciation, and the domestic target language, the learning environment is far from ideal, with few competent teachers. In addition, such learning cannot achieve computer-assisted language learning (CALL) technology. The efficient combination of computer technology and language teaching and learning methods provides a new solution to this problem. The core of CALL is speech recognition (SR) technology and speech evaluation technology. The development of deep learning (DL) has greatly promoted the development of speech recognition. The pronunciation resource collected from the Chinese college students, whose majors are language education or who are planning to obtain better pronunciation, shall be the research object of this paper. The study applies deep learning to the standard but of target language pronunciation and builds a standard evaluation model of pronunciation teaching based on the deep belief network (DBN). On this basis, this work improves the traditional pronunciation quality evaluation method, comprehensively considers intonation, speaking speed, rhythm, intonation, and other multi-parameter indicators and their weights, and establishes a reasonable and efficient pronunciation model. The systematic research results show that this article has theoretical and practical value in the field of phonetics education.

1. Introduction

Speech processing technology has received considerable attention recently due to a variety of applications in the areas of automated speech recognition, information retrieval, and assisted communication. There has been a lot of research on voice processing for different human languages all around the world. As a result, deep learning has become increasingly capable of independent speech processing, including voice recognition and synthesis, in recent years.

With the progress of globalization, common langauges such as Chinese or English for international communication as a widely used language have attracted more and more people's attention. Hence, a common language is required for academic research exchanges, and it is also required for industrial production and programming to view technical documents [1–3]. Because China wishes to successfully promote internationalism and keep lines of communication open for trade and business matters, English used today or moreover, Chinese, both used in the future has become extremely vital. As one of the world's great powers, it is in the country's best interests to stay abreast of changing worldwide trends. Therefore, an international language is becoming more and more important for the people who are eager to communicate with another around the world. For Chinese, mute English has always been the number one problem in learning English, and speaking fluent English is the dream of many Chinese people [4] as well as the dream of speaking normal mandarin by non-Chinese speaking people. This paper aims at alleviating the problem of relatively low accuracy of oral pronunciation by using the DL algorithm to provide feedback on speech evaluation. The academic research on the automatic scoring algorithm can be traced back to the early 1990s. At present, the mainstream spoken-reading scoring method adopted by the

industry is mainly an SR engine based on the hidden Markov model. We use its likelihood score value and other relevant information as the basis for scoring. The advancements in DL, big data, and cloud computing have also had an impact on voice recognition and evaluation technology [5]. DL is taken from the discipline of machine learning and tries to create and imitate the human brain's deep neural network (DNN) for analysis and learning purposes [6]. DNN has proven to be a powerful tool for resolving a variety of challenging issues. Due to its ability to better mimic human brain neurons and execute multi-layer deep transmission, it has been shown in the field of SR. Research on fundamental DL in English SR technology can considerably enhance the ability of speech information processing, increase user efficiency, and improve user experience [7]. At home, many people utilize portable devices like language repeaters and cell phones to help them learn oral langauge pronunciation; however, these tools cannot conduct voice recognition directly and are limited instead in their capabilities. Evaluate students' pronunciation and feedback objectively and accurately. In addition, several CALL systems at home and abroad primarily focus on the learning of vocabulary and grammar because of the limitations of technology. Only one or two assessment indications are used as the foundation for evaluation, and there are functional problems that can only offer students an overall score [8]. Despite this, the traditional evaluation of pronunciation in oral language teaching still relies on manual scoring with considerable subjective willingness, diverse criteria, and a slow pace. The rating of the same pronunciation will be affected by the differing knowledge and experience of scoring experts, as well as the different standing of the same expert. Subjective characteristics such as these contribute to the lack of consistency and stability in manual evaluation. Additionally, manual grading will need a significant amount of time and money [9].

This paper is organized such that Section 2 defines some related work. Section 3 proposes the main methods of the study. Section 3 also explains the DL neural network, the basic idea and training process, restricted Boltzmann machine, deep belief network, and multi-parameter pronunciation quality evaluation. Section 4 defines the experiment and analysis of the proposed work. Finally, the paper ends with a conclusion in Section 5.

2. Related Work

DL is a type of machine learning algorithm that attempts to obtain a high-level abstract representation of data that includes multiple layers of non-linear mapping [10–12]. DL is within the scope of representation learning in machine learning. Data can be expressed on many levels. For example, an image can be expressed as individual pixels or more abstract corner features, some of which are more conducive to specific tasks. One of the most commonly used scenarios of DL is to use unsupervised or semi-supervised algorithms to automatically learn features to replace manually designed features. DL attempts to learn a better representation of data from large-scale unlabeled data, so DL is also called representation learning or unsupervised feature learning algorithms. For example, DNN, convolutional deep neural network (CDNN), and DBN have been widely used in computer vision, automatic SR, natural language processing, and other fields with unprecedented success.

The DL network structure started from the perceptron-based multi-layer artificial neural network model (ANN) introduced in [13]. The history of ANN is even longer. In 1989, Professor Yann successfully applied the standard BP algorithm to a deep ANN model for the first time. This ANN model was used for the recognition of handwritten zip codes in the American mail system [14]. In addition, the multi-layer neural network was not successful in a wide range at that era due to various reasons. A major factor is the problem of gradient dispersion under the multi-layer neural network. The work related to DL achieved a breakthrough around 2006. Professor Hinton demonstrated how to train an unsupervised Boltzmann machine hierarchically and then use the BP algorithm to fine-tune the parameters of these stacked Boltzmann machines. A DL network with excellent performance can be efficiently trained [15, 16]. After this resurgence, DL has achieved the best results in many different fields, especially in the fields of computer vision and SR. In the database classification experiments represented by TIMIT and MNIST, the effect of DL algorithms is the best, and the DL structure with CNN as the core structure performs well [17–20]. The great influence of DL in the industry began with the application of DL in the field of large-scale SR. At the end of 2009, Geoff Hinton and Li Deng jointly organized the 2009 NIPS (Neural Information Processing Systems) seminar on the application of DL in SR. The purpose of this seminar is to discuss the limitations of deep generative models in the field of SR and the possibility of the development of DL in the future high-performance computing big data era. This seminar believes that pre-training with fine-tuning of parameters is the main method to prevent gradient dispersion in DL. However, not long after this seminar, Microsoft research found that when there is a large amount of training data, only a reasonable design of the DNN structure is needed, and a DL algorithm with better performance than GMM-HMM can be obtained without pre-training [21]. The improvement of hardware performance is of great help to the success of DL, especially the use of high-performance GPUs (graphics processing units) for DL training, which has greatly promoted the training time of DL from several weeks to several days [22]. The algorithm uses an extreme learning machine as a weak classifier, and it uses the "AdaBoost" framework to improve the effectiveness of the classifier. The finally obtained classifier overcomes the extreme learning machine due to partial weights given randomly. The shortcoming of instability is compensated and the classification accuracy is higher. The "AdaBoost" algorithm is an integrated algorithm, proposed in reference [23]. It has the advantages of efficient and fast training and is not easy to overfit.

All in all, the current concept of DL is very hot, providing high-accuracy and high-speed calculations for SR and creating new opportunities for intelligent voice interaction. We are in the era of big data. AI algorithms with DL at the core are improving human life in all aspects. A large number of Internet resources make it easier and easier to obtain data. With the development of science and technology, computers

are becoming cheaper and cheaper. The performance of the computer is getting better and better. The abundance of computing resources and the convenience of data acquisition have enabled more and more algorithms to achieve better performance. Corresponding to the convenience of data acquisition, the acquisition of class labels is relatively difficult. Therefore, unsupervised feature representation algorithms that do not require class labels have more and more application value. With the increase in the level of data, simple laboratory research has become more and more inadequate. More and more scientists have joined the industry, complementing the advantages of enterprises, taking into account scientific research and industrial reality, academic guidance industry, and industry feeding back academics. This change also deeply affects the development of society. In this era of big data, there is still much we can do.

3. Method

In this section, we define DL neural network, the basic idea and training process, restricted Boltzmann machine, deep belief network, and multi-parameter pronunciation quality evaluation.

3.1. DL Neural Network. DL is a sort of unsupervised learning derived from ANN and a multi-layer perceptron with numerous hidden layers (HLs). Using unsupervised data, Hinton devised a DL algorithm for neural networks, allowing for the training of neural networks with at least seven layers, or what is known as a DNN. To identify dispersed feature representations of data, DL uses a combination of low-level features to create a high-level, abstract representation. Because of its emphasis on large-scale training data and multi-layered machine learning models, DL aims to improve prediction or classification accuracy by uncovering the most fundamental aspects of the input data. "Feature learning" is the goal, while "deep model" is the methodology. DL differs from shallow learning; in that, it highlights the depth of model structure. The number of concealed nodes might range from five to ten tiers. With fewer parameters, multi-layer expresses greater complexity. We highlight the importance and necessity of feature learning, which is the process of transforming an original feature space into new feature space in order to discover a distributed feature representation of the data, which is more easily predicted or classified than an original feature representation. The process of artificially creating features can better represent the data's key information when using large data to learn features. DL technology is currently being used to solve the problem of classifying many patterns, and its effectiveness has been demonstrated in SR. As a relatively new area of study, it is certain to have a big impact on AI and machine learning. DL is used to increase the accuracy and speed of the studied language's SR in this article.

3.2. The Basic Idea and Training Process. There are three main principles that underlie DL, all of which revolve around the use of unsupervised and supervised learning to train each layer of the network and then use the outcomes of those layers to alter all other layers.

The approaches are as follows:

(1) Construct single-layer neurons layer by layer and train only one single-layer network at a time.

(2) When all layers have been trained, apply the wake-sleep algorithm to fine-tune parameters.

This results in all layers except the top layer, which remains a single-layer neural network, becoming graph models. Cognitive functions are carried out with the help of the upward weight, while information is generated with the help of the downward weight. Create more or less equal weights for all of the variables using the wake-sleep algorithm after that. Maintain as much agreement as feasible between cognition and generation, which means the top-level representation of the generation should be able to restore the bottom nodes with as much accuracy as possible. One of the two elements of the wake-sleep algorithm is that of wake and sleep: by using gradient descent and external features to produce an abstract representation of each layer and adjusting the weight across layers if the reality is not the same as the imagination, the weight changing can make this imaginaiton a real one. When a person is asleep, the generation process generates the state of the bottom layer and simultaneously alters the weights between layers.

The training process of DL is as follows:

(1) Use bottom-up unsupervised learning. Unsupervised stratification training for parameters of each layer was carried out by using uncalibrated data. The most significant distinction from typical neural networks is that the phase is identical to the feature learning procedure. When training a neural network, the first layer is trained with uncalibrated data, and these data are used to learn the parameters of that layer. After training $n-1$, the output of $n-1$ is used as an nth input for the n-1st layer, the nth layer is trained, and nth layer parameters are produced for each layer.

(2) A top-down supervised learning style is conducted. In the first step, supervised learning is used to further tune the parameters of the overall multi-layer model based on the parameters of each layer. There are many advantages to using DL instead of traditional neural networks, such as better results because the DL initialization parameters are learned rather than randomly initialized. This means that the DL initial value is closer to a global optimal value, which means better results can be achieved. Therefore, the first phase of feature learning is responsible for the majority of DL's impact.

3.3. Restricted Boltzmann Machine

3.3.1. Overview of RBM. RBM is composed of two layers: one that is visible and one that is hidden. It is composed of some visible units and some hidden units. Both visible and hidden variables are binary variables, that is, the entire

network is a bipartite graph, with full connections between layers and no connections within the layers. In other words, there are edges only between visible and hidden units, and there is no edge connection between visible and hidden units. The specific model is shown in Figure 1.

3.3.2. RBM Learning Algorithm. The method of describing RBM is the energy function (EF) and the PD function. Combining the two means that the PD is a functional of the EF. To express the joint configuration's energy for both the visible and unobservable components, the formulas are used:

Here, the parameters of the RBM model represent the deviation of VL node and the deviation of HL node and represent the connection weight between VL node and HL node.

Based on the EF, the joint PD of a certain configuration can be determined by the Boltzmann distribution and the energy of the configuration.

The normalization factor (NF) is also known as the partition function.

Because the HL nodes are conditionally independent, that is:

Further, by factoring the above formula, we can get the probability that the j^{th} node of the HL is 1 or 0 on the basis of a given VL:

In the same way, on the basis of a given HL, the probability that the i^{th} node of the VL is 1 or 0 can be obtained as

After a given training sample, training an RBM means learning to adjust the parameter to fit the given training sample, and even under this parameter, the PD represented by the corresponding RBM must match the training data as much as possible.

3.3.3. RBM Evaluation Method. For an RBM that has been learned or is learning, the simplest evaluation index is the log-likelihood of the RBM to the training data. However, due to the existence of the NF, the computational complexity is quite high, so only sampling approximate methods can be used to evaluate the pros and cons of RBM. The commonly used approximation method is reconstruction error, which is the difference between the original data after a Gibbs transfer through the distribution of RBM with the training sample as the initial state. The reconstruction error can evaluate the likelihood of the RBM to the training samples to a certain extent, but the reliability is not high. But in general, its calculation is quite simple and the overhead is small, so it still has considerable value in practice.

3.4. Deep Belief Network. In 2006, Geoffrey Hinton proposed the DBN. Layer-by-layer, DBN uses an unsupervised and greedy algorithm. The weights of the created model are pre-trained using this method, and then the backpropagation algorithm is used to fine-tune the network to produce a better model. DBNs with appropriate configurations have been found to be superior to random initialization when it comes to setting up multi-layer perceptrons. The deep Boltzmann machine (DBM) can be obtained by increasing the number of HLs and using the unsupervised greedy layer-by-layer method. By training RBM layer by layer, DBN achieves a globally optimal initial parameter that improves network performance. According to numerous studies, a significant number of labeled training sets are required, the convergence speed is slow, and the incorrect parameter selection causes the network to fall into a local optimum when using a DBN instead of a typical BPN.

3.5. Multi-Parameter Pronunciation Quality Evaluation

3.5.1. Evaluation Index. The prosody of pronunciation is a very important factor in sentences. Each language has its own characteristics in terms of prosody, and sentences that cannot grasp the prosody of the language will appear unnatural. The evaluation of pronunciation quality is mainly to comprehensively evaluate the pronunciation standard, length, and rhythm of the speech. To evaluate pronunciation, phonemes and words are mostly judged on their conventional pronunciation, whereas sentences or paragraphs are judged on their prosodic features, which are a major factor in how they convey their meaning. When evaluating the pronunciation quality of sentences and paragraphs, comprehensive consideration should be given to prosodic information such as whether the speaker can more accurately grasp the key information of the sentence, whether the less important information of the sentence is relatively weak, and whether the sound length is appropriate. Therefore, this article uses two major indicators of pronunciation standard and prosody to evaluate pronunciation quality. From the perspective of linguistics, prosody, or rhythm, refers to phonological arrangement above the segment level. In a way, it is an organized way of putting together distinct linguistic components into conversation or discourse blocks. In addition to conveying linguistic information, prosody can also convey paralinguistic and non-linguistic information, as well as express mood and attitude, identify vocabulary, and perform other roles. Suchpronunciation quality can be perceived to be closely linked to its rhythm. Super-segment characteristics, or prosodic features, refer to the speed, rhythm, and intonation created by dynamic patterns of associated factors like pitch, intensity, and duration. Evaluation of the pronunciation and feedback instructions are provided using speech speed, rhythm, and pitch to assess the quality of the language's pronunciation, as well as the distinct energy and pairwise variation index (DPVI).

3.5.2. Pronunciation Standard Evaluation. If the content material is correct and full and pronunciation is fluid and clear, the standard evaluation checks whether there are any pronunciation errors. Pronunciation standards are evaluated using the MFCC coefficients based on the human auditory model, and the SR model is built using the DBN to determine whether the material is complete and correct. For example, Figure 2 illustrates how MFCC characteristics can be used in conjunction with the correlation coefficient of the standard sentence and the MFCC feature of an input

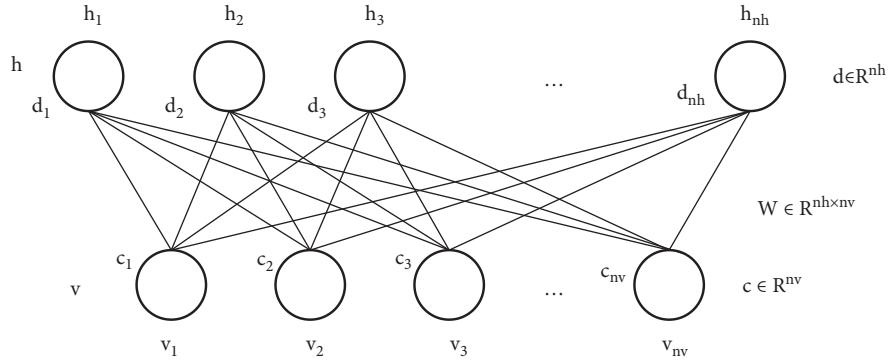

Figure 1: RBM model.

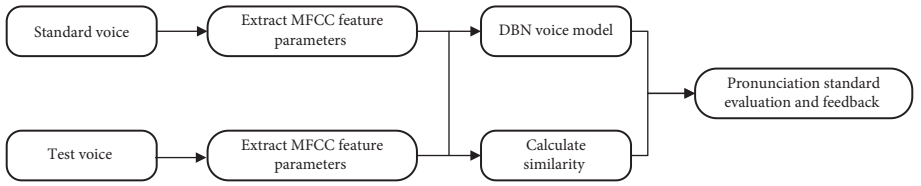

Figure 2: Pronunciation standard evaluation process.

sentence to identify whether or not the pronunciation is clear and fluent, as illustrated in the figure.

3.5.3. Speaking Rate Evaluation. It is common for speech speed to be measured in terms of pronunciation speed, which is a measure of the speaker's ability to pronounce words quickly. By counting how many syllables are pronounced each second in a unit of time (T), it can be approximated as the entire speech time, including pauses (N). Depending on how fast a speaker is speaking, the pronunciation of a single statement might vary greatly from person to person. In addition, the speaker's emotional condition will also influence the speed of speaking. As an example, the speed of speech tends to be faster in furious and pleased states than in a calm one, while it is slower in sad situations. This article adopts speech rate evaluation based on speech duration and calculates the ratio of test sentences and standard sentences, as shown in the following formula.

3.5.4. Rhythm Evaluation. The rhythm of the language is the similarities and differences in height, severity, length, and priority of the phonetics, and it appears regularly and alternately with certain types of phonetic unit fragments. Rhythm is divided into three types: completely accented, incomplete accented, and emphasized accented. When reading and speaking, the rhythm groups formed by different combinations appear alternately, and its meaning function is to enhance the melody and musical sense. The research on the rhythm of language has been relatively mature. The hypothesis of temporal synchronization of language rhythm defines language rhythm as the isochronous repetition of a certain language unit segment, and according to the isochronous characteristics of speech, language is divided into stress-timed and syllable timing language. Languages like Chinese, French, Italian, and Spanish use syllable timing; therefore, the intervals between subsequent syllables are essentially the same in each language. Not like Chinese, English is a typical stress-based language, that is, the basis and main body of each sentence in English are stressed syllables, and the number of stressed syllables determines the beat of the sentence.

The sentences in the studied language have the following three characteristics:

(1) Generally speaking, the higher the frequency of stressed syllables in the sentence is, the slower the speaking rate is and the clearer the syllables will sound.

(2) The unstressed syllables appear crowded among the stressed syllables. It seems brisk and vague.

(3) The length of time needed to speak a sentence does not depend on the number of words or syllables in the sentence, but more importantly, it depends on the number of stressed syllables in the sentence.

Stressed syllables play a role of emphasis and contrast in sentence organization and semantic expression and have the following three characteristics:

(1) Loud.

(2) Long pronunciation.

(3) Clear and easy to distinguish.

The rhythm evaluation mechanism is shown in Figure 3.

3.5.5. Intonation Evaluation. Intonation refers to the configuration and change of vocal tone. Intonation changes according to different modes in the unit of the meaning group, and its meaning function is manifested in expressing a variety of different emotional colors. Various intonations

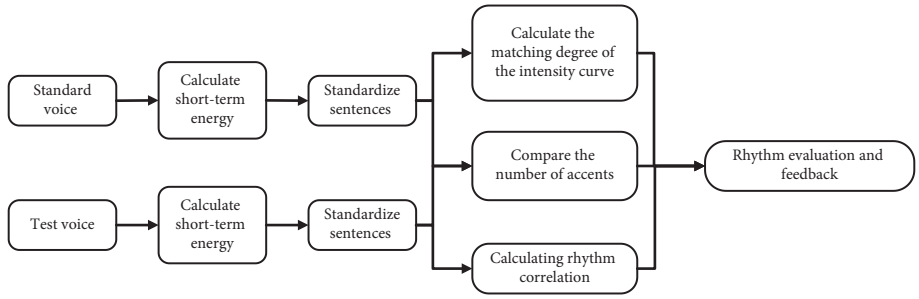

Figure 3: Rhythm evaluation and feedback process.

of the same speech might convey different meanings. Three primary intonations: rising, falling, and flat, are used in the language of English. Intonation evaluation aims to use calculations to automatically determine whether the pronunciation intonation is standard or not and to prompt the difference between it and the standard voice. To understand intonation, the pitch is the most fundamental and crucial aspect. The pitch of a sound is directly related to the number of vibrations: the more the vibrations, the higher the sound; the fewer the vibrations, the lower the sound. Pitch is a measure of how quickly the sound changes in pitch. The vocal cords' fundamental frequency change is what gives speech its material form of pitch. From the change of the fundamental frequency, the different modes of intonation change can be determined, that is, the pitch can determine the different modes of intonation. Traditionally, pitch tracking methods are separated into two categories: time-domain method and frequency-domain method. Time-domain approaches primarily include autocorrelation function (ACF), average magnitude difference function (AMDF), and simple inverse filter tracking (SIFT). The harmonic product spectrum method and cepstrum method are the two most commonly used techniques in the frequency domain.

In these methods, the basic process of pitch tracking is as follows:

(1) Framing the entire speech signal.

(2) Calculating the pitch corresponding to each speech frame.

(3) Excluding unstable pitch values, you can set the pitch filter in the high-value range.

(4) A common method is to use median filters to smooth out the entire pitch.

Frames of voice data are first segmented, and then the following analysis is done in frames. Each frame of data associated with an English sentence is extracted using ACF in the time domain, which is further used in the range of pitch values to exclude unstable and abnormal voice frames, before using a median filter to smooth out all pitch, and finally using the DTW algorithm to determine how well a standard sentence and the input sentence fit intonation-wise.

4. Experiment and Analysis

In this section, we describe the data sources, SR experiment, and voice evaluation experiment in detail.

4.1. Data Sources

4.1.1. Spoken Arabic Digit Dataset. This paper draws on the "Arabic Phonetic Numerals" data set from the Machine Learning library at the University of California, Irvine (UCI). Following the extraction of 13-order MFCC feature parameters, this dataset contains the pronunciation of Arabic numerals, which includes a total of 8800 voice samples. Before extracting the MFCC feature parameters, the parameters that need to be set are the sampling rate of 16 KHz, the 16 bit encoding, the hamming windowing function, and the pre-emphasis filter function ($1–0.96Z^{-1}$).

4.1.2. Data Sources of the Sentences. The subjects in this article are randomly surveyed college students, a total of 30 people, including 15 boys and 15 girls. The subjects used recording software to record with a sampling rate of 16 kHz and 16 bit encoding.

There are 10 recorded sentences, which are common or classic sentences in the spoken expression of the language studied, as shown below:

(1) Nothing is difficult if you put your heart into it.

(2) Once upon a time there were six blind men who lived in a village in India.

(3) I have butterflies in my stomach.

(4) Mind your own business.

(5) Take my word for it.

(6) Falling in love with yourself first does not make you vain, it makes you indestructible.

(7) It never rains but it pours.

(8) Whatever is worth doing is worth doing well.

(9) It is better to be alone than to be with someone you are not happy to be with.

(10) Happiness is a way station between too much and too little.

4.2. SR Experiment

4.2.1. Data Preprocessing. For the purpose of determining the efficacy of the model presented in this research, the recognition rates of this model and other models under the recognition of isolated words of unnamed people are compared in a series of experiments. We use the dataset containing a total of 8800 Arabic digit speech data (88 people's pronunciation of 10 Arabic numbers, where each number is repeated ten times), with 6000 pronunciations of the first 60 people used as a training set and 2000 pronunciations of the last 20 people used as a test set (a total of 8800 Arabic digit speech data). For neural networks, most of them have time regulation problems. It is necessary to translate variable-length speech feature parameters into feature vectors with equal length since the structure of the neural network classifier is fixed, and speech feature parameters have the problem of unequal dimensionality, which is why we need to transform them. In this paper, a piecewise average method is used to perform preprocessing operations such as dimension reduction and regularization on the speech feature parameters of the "Spoken Arabic Digit" dataset. First, the speech signal feature parameters are averagely segmented. The speech feature parameters can be expressed as, where the order of the feature parameter is, the number of frames of the feature parameter after segmentation is, and the number of original speech frames is. Then, the calculation formula for dividing the characteristic parameters into segments is as follows. Where is the speech feature parameter of the segment after segmentation.

After that, the averaging operation is performed on the frame parameters of each sub-segment, and the mean vector of each sub-segment is obtained. Finally, after collecting the mean vector for each sub-segment and combining the mean values into a matrix, a matrix of size has been obtained and is the characteristic parameter output value after dimensionality reduction and regularization. Table 1 shows the processes of dimensionality reduction and regularization of voice feature parameters.

From the data in Table 1, we can see that the segmented mean dimensionality reduction and regularization algorithm can reduce the dimension of the characteristic parameter matrix of size into a parameter matrix of size. It can be seen from the formula that the segmented mean dimensionality reduction and regularization algorithm successfully removes the influence of the number of speech frames on the data size after dimensionality reduction and regularization. The segment size is related to the sub-segment size so that speakers of different lengths can be regularized into a matrix of the same size, which greatly facilitates the implementation of the SR algorithm and greatly improves the performance of the SR algorithm. In the experiment in this article, each voice signal is normalized to 16 frames, that is, the data volume of each voice signal is $13 * 16 = 208$. Furthermore, normalize the "Spoken Arabic Digit" dataset after dimensionality reduction to [0, 1], which is beneficial to reduce the impact of feature differences caused by differences in speakers and channels. In addition, the DNN uses small batch data processing to reconstruct error adjustment weights, making the network more stable and operating efficiency higher. The dataset has a certain order, so it is necessary to randomly scramble the dataset before batch processing to improve the performance of the model.

4.2.2. Experimental Results and Analysis. At this time, there is no unified standard for the selection of related parameters such as the number of HLs in the DBN model, the number of nodes in each HL, the number of samples processed in batches, and the number of iterations of the BP network, among other things. Because of this, all of the parameters of the DBN model built in this article were determined through experimental comparison and tuning. This paper's DBN model was developed after a large number of trials and has four layers: an input layer, two HLs, and an output layer, among others. Specifically, the number of nodes in each layer is $208 * 1000 * 1000 * 10$, where the number of nodes in the input layer corresponds to the amount of feature parameter data contained in the input voice, and the number of nodes in the output layer corresponds to the number of output categories contained in the output voice. In the process of using the BP network to adjust the weight of the error backpropagation of the DBN model, appropriately reduce the number of samples in each batch and increase the number of iterations. Iteratively training a small number of times can improve the recognition rate of the model. Of course, it is also accompanied by the growth of model building time. It is worth noting that neither the lower the number of samples are in the batch, the more the iterations are, nor the lower the number of samples are in the batch, the higher the recognition rate is. Moreover, it is not a monotonic linear relationship. Too few batch samples and too many iterations may also cause the model to overfit and reduce its performance of the model. For the "Spoken Arabic Digit" dataset in the same UCI machine learning library, Hammami et al. proposed a tree distribution approximation based on graphical tree structure (TDA-GTS); compared with tree distribution approximation based on maximum weight spanning tree (TDA-MWST), traditional discrete hidden Markov model (DHMM), and continuous density hidden Markov model (CDHMM), the recognition effect has been improved. Huang proposed a K-means clustering algorithm based on selective weights and thresholds (KASWT) and compared it with the BP AdaBoost algorithm; the recognition effect has been improved. Here, the model in this paper is compared with the above models, and the comparison results of the recognition rate are shown in Figure 4.

It can be seen from Figure 4 that the recognition rate of the DBN model constructed in this paper is 97%, which is better than the above models, so it is reasonable and effective.

4.3. Voice Evaluation Experiment. The goal of this experiment is to test the effectiveness of the model and approach provided in this article for evaluating the quality of the pronunciation by comparing the results from the machine

Table 1: Speech feature parameter dimensionality reduction and normalization.

Parameter	Stage					
	1	2	3	4	5	6
Matrix size number						

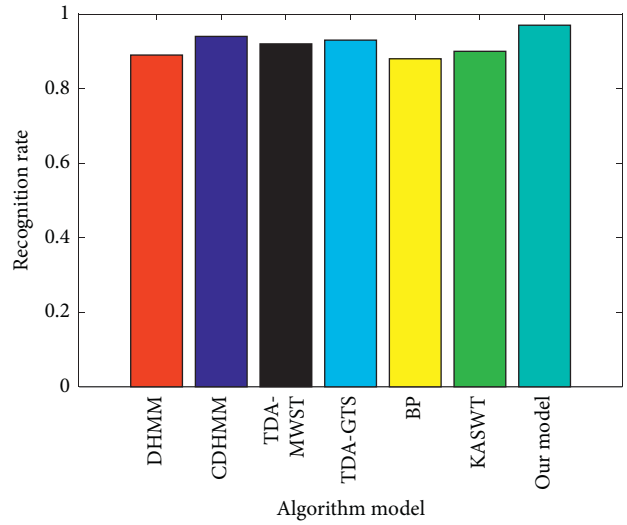

Figure 4: Comparison of recognition rates under different models.

and human evaluations of identical sentences. There are two steps in this process: checking for manual reliability and then checking for consistency between machine and manual dependability based on the manual evaluation's credibility. The consistency of machine evaluation and manual evaluation is expressed using agreement, adjacent agreement, and the Pearson correlation coefficient in this paper.

4.3.1. Manual Evaluation. In accordance with the suggestions of relevant phonetics experts, we have set four different levels of 1, 2, 3, and 4 for different evaluation indicators (pronunciation standard, speaking speed, rhythm, and intonation) and overall evaluation based on the quality of pronunciation characteristics of college students with various oral phonetic proficiency. As can be seen in Table 2, the various evaluation levels and their accompanying evaluation criteria are listed.

The manual evaluation was completed by 5 college phonetic teachers with rich teaching experience. They, respectively, evaluated 10 common sentences in recorded spoken fragments of 30 randomly selected college students, including 4 evaluation indicators of pronunciation standard, speaking speed, rhythm and intonation, and the overall evaluation situation. Considering that the subjectivity of the teacher in the manual evaluation process may affect the evaluation results, this paper uses the Pearson correlation coefficient to test the reliability of the manual evaluation results. For the convenience of calculation, the evaluation grades I, II, III, and IV are converted into corresponding scores of 4, 3, 2, and 1, respectively. Pearson correlation analysis shows that the scores of the four evaluation indicators, namely, pronunciation standard, speech speed, rhythm, and intonation, or the total score is positively correlated. This further illustrates that the five teachers maintained basically the same evaluation standards during the evaluation process, which effectively guarantees the reliability of the experimental data. Furthermore, the evaluation results of 5 teachers are averaged (rounded up), and the evaluation indicators and overall scores of different sentences of different students are obtained as the final manual evaluation result.

4.3.2. Experimental Results and Analysis. According to the method introduced in this paper, 30 students were investigated. A total of 300 sentences can be obtained by sampling 10 sentences from each student. Then, the 300 sentences are evaluated according to four evaluation criteria, namely, pronunciation standard, speaking speed, rhythm and intonation, and the corresponding scores are obtained, which shall be compared with the scores obtained by manual evaluation. The experimental results are shown in Figures 5–8.

4.3.3. Test of the Evaluation Model. This article takes into account pronunciation standards, speech speed, rhythm and intonation, and other multi-parameter indicators and their weights, as well as regression analysis, in order to construct a reasonable and objective pronunciation quality evaluation model for college students on the basis of testing the credibility of the aforementioned evaluation indicators. By employing mathematical statistics to develop statistical models, we can uncover statistical relationships between objective variables, and by collecting a large number of experiments and observation data on objective things, we can search for statistical regularities hidden within seemingly unpredictable phenomena and make model predictions based on the models developed. The overall score of the manual evaluation is utilized as the dependent variable in this article, while the scores on the pronunciation standard, speaking speed, rhythm, and intonation are used as the independent factors in this study. Selective attention is paid to the sentences that are completely consistent with both manual evaluation and machine evaluation. Multiple linear regression analysis is used to obtain each evaluation using SPSS software. The weight of the index is expressed as follows.

PSS stands for pronunciation standard score, SPS stands for speech speed score, RHS stands for rhythm score, and INS stands for intonation score.

The F test and the t-test are two statistical test methods used to determine the significance of the regression equation and its associated regression coefficient, respectively. SPS, RHS, and INS have significant effects on the random variable score, and hence the multiple linear regression equation's significance can be tested using the F test. SPSS program creates the variance analysis table, which indicates a significant linear correlation between PSS, SPS, RHS, and the INS, i.e., the four assessment indicators collectively have a significant influence. Only 0.005 percent of the time is it possible to make a mistake that has a substantial linear effect. The t-test is used to determine the significance of the

Table 2: Related evaluation indicators and their grading standards.

Level	Pronunciation standard	Speaking speed	Rhythm	Intonation	Total
1	Complete and accurate	Moderate	Strong rhythm	Accurate	Excellent
2	Almost complete and accurate	Slightly fast	Good rhythm	Almost accurate	Good
3	With a little error	Fast	Mediocre rhythm	Basically accurate	Pass
4	Incomplete with serious errors	Too fast	Poor rhythm	Inaccurate	Poor

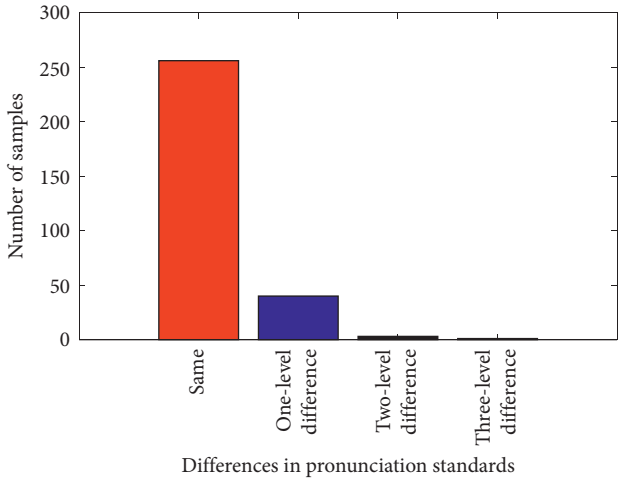

Figure 5: Comparison of sample number of pronunciation standard difference grade.

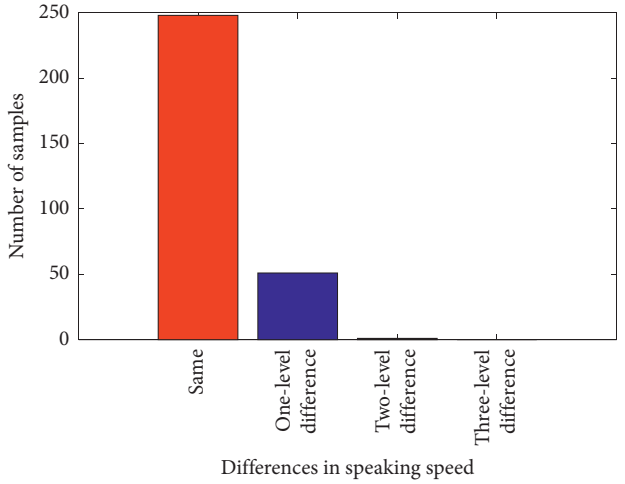

Figure 6: Comparison of sample number of speaking speed difference grade.

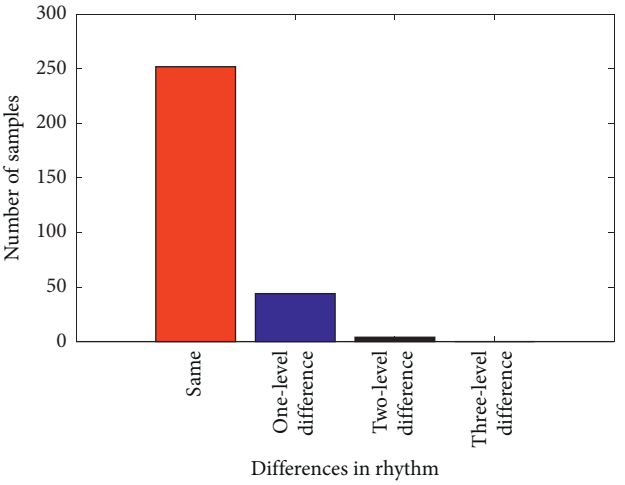

Figure 7: Comparison of sample number of rhythm difference grade.

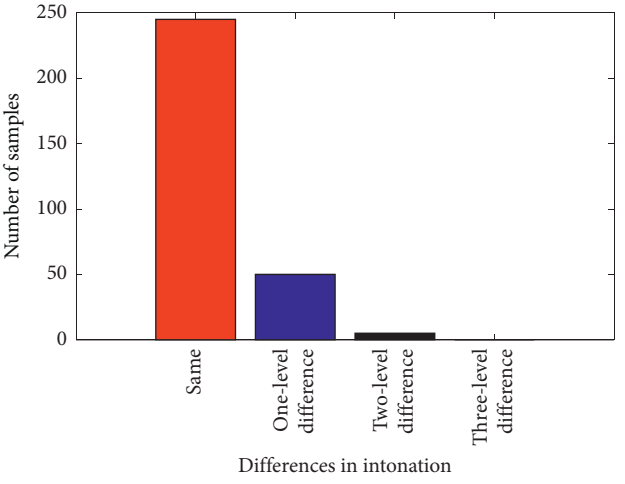

Figure 8: Comparison of sample number of intonation difference grade.

regression coefficient in regression analysis. The SPSS regression coefficient table has been made public. There are four evaluation indicators that all pass significance tests at a significance level of 0.05. Additionally, this article employs a formula to evaluate the performance of 30 pupils, 10 sentences, and 300 sentences. Table 3 displays the findings of the research. Confusion rates between machine and manual assessments are 88.55%, with an adjacent coincidence rate of 100% and a Pearson correlation coefficient of 0.751, which indicates a significant correlation between the two methods.

The experimental results are very good, but there are still some deviations, which may be related to the following factors:

(1) The amount of data in this article is too small, and linear regression analysis is used to fit the model.

Table 3: Overall evaluation of experimental results.

Index	Difference		
	Agreement rate (%)	Adjacent agreement rate (%)	Pearson
Pronunciation standard	88.25	100	0.751

(2) This article uses 5 phonetic teachers' average scores as the final score of the manual evaluation, and the rounding method makes the score high.

(3) According to the four-level scoring method and the adjacent coincidence rate, the difference between the machine and manual assessments multiplied by one level for each of the four levels is used to calculate it.

To a certain extent, the adjacent coincidence rate is increased. According to the previous definition, the pronunciation standard has the largest weight, the rhythm and intonation have the second weight, and the speech speed has the smallest weight. After obtaining the above experimental results, we communicated with phonetics experts to further investigate the reliability of the multi-parameter pronunciation quality evaluation model proposed in this article. According to experts, when evaluating the pronunciation quality, the most important indicators are pronunciation standards, which require accurate content, fluent pronunciation, and no obvious pronunciation errors; rhythm and intonation primarily express the emotional color of the speaker, enhance the melody of the voice, and allow the voice to be closer to the reality of life; speech speed varies from person to person, as long as it is not too fast or too slow to affect the quality of the pronunciation, and the speed of speech does not affect the quality of the pronunciation. Therefore, the result is reasonable. In summary, the standard evaluation model of pronunciation in teaching for college students in this article is credible.

5. Conclusion

Technology for voice recognition and evaluation is crucial to computer-assisted speech learning. One of the most important technologies is voice recognition technology, which is essential and plays a major role. As a result, SR is an essential foundation and prerequisite for a speech evaluation, and only high-accuracy SR can further achieve good speech evaluation findings. This paper examines the current issues in voice recognition technology by examining the advantages and limitations of the traditional SR algorithms DTW, HMM, and ANN. In spite of their accomplishments thus far, they have met unprecedented blockages and it is impossible for them to further improve their accuracy and speed. Neural networks can be reinvigorated with the use of DL technology, and SR technology is also being improved on a daily basis. Using a multi-layer non-linear neural network structure, the DNN can express complex functions with fewer parameters and better find the distributed properties of the data through an unsupervised, layer-by-layer feature transformation, demonstrating superior feature learning capabilities. In addition, it aids in the improvement of categorization or prediction accuracy. DL technology is used for phonetic recognition, and an SR model is built using DBN. The recognition results are better than the enhanced hidden Markov model, BP neural network model, and tree distribution approximation model based on the "Spoken Arabic Digit" dataset in the UCI machine learning library. In addition, the DBN model is used to evaluate the quality of the pronunciation, and the SR model based on the DBN is employed in the evaluation of the pronunciation standard.

Computer-aided language learning systems at home and abroad tend to emphasize the acquisition of vocabulary and grammar rather than the development of speech and pronunciation. Only an overall score can be assigned to a learner's pronunciation because there are only a few evaluation indications and certain functional flaws. In the aspect of oral pronunciation evaluation, oral pronunciation test is still dominated by manual scoring, which is highly subjective, slow, and poor in standardization, repeatability, and stability.

Data Availability

The datasets used during the current study are available from the corresponding author on reasonable request.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This study was supported by the Educational Science Project of Jilin Province (Research on the Development Path of Preparatory Education in China under the Background of Big Data) (project no. 1805313). This study was also supported by the Education Department of Jilin Province (Research on Construction and Application of Second Language Intonation Model Based on Spoken Chinese Output) (project no. 2005209).

References

[1] H. Abubakar Muhammad, S. Ya, and U. Aliyu, "Teaching and learning English language In Nigerian schools: importance and challenges," *Teacher Education and Curriculum Studies*, vol. 3, no. 1, pp. 10–13, 2018.

[2] E. Amiri and L. Branch, "A study of the application of digital technologies in teaching and learning English language and literature[J]," *International Journal of Scientific & Technology Research*, vol. 1, no. 5, pp. 103–107, 2012.

[3] C. Li, "Study of mother tongue negative transfer on senior high school students' English grammar learning-A case study of the middle school of jincheng mining in shanxi Province [J]," *Journal of Frontiers in Educational Research*, vol. 1, no. 1, pp. 46–61, 2021.

[4] Z. Fu, C. Ji, H. Weiss-Krumm, G. Wang, and Y. Ma, "Chinese-to-English phonetic transfer of Chinese university EFL students[J]," *Asian Journal of Applied Linguistics*, vol. 7, no. 1, pp. 18–31, 2020.

[5] V. Z. Këpuska and T. B. Klein, "A novel Wake-Up-Word speech recognition system, Wake-Up-Word recognition task, technology and evaluation," *Nonlinear Analysis: Theory, Methods & Applications*, vol. 71, no. 12, pp. e2772–e2789, 2009.

[6] S. Han, H. Mao, and W. Dally, "Deep compression: compressing deep neural networks with pruning, trained quantization and huffman coding[J]," *Fibers*, vol. 56, no. 4, pp. 3–7, 2015.

[7] G. Dahl, D. Yu, L. Deng, and A. Acero, "Context-dependent pre-trained deep neural networks for large-vocabulary SR[J]," *IEEE Transactions on Audio Speech and Language Processing*, vol. 20, no. 1, pp. 30–42, 2011.

[8] T. Elliott, "Memory compression and thermal efficiency of quantum implementations of nondeterministic hidden Markov models[J]," *Physical Review A*, vol. 103, no. 5, Article ID 052615, 2021.

[9] Y. Yang and Y. Yue, "English speech sound improvement system based on deep learning from signal processing to semantic recognition," *International Journal of Speech Technology*, vol. 23, no. 3, pp. 505–515, 2020.

[10] C. Tejedor-García, D. Escudero-Mancebo, E. Cámara-Arenas, and C. González-Ferreras, "Assessing pronunciation improvement in students of English using a controlled computer-assisted pronunciation tool[J]," *IEEE Transactions on Learning Technologies*, vol. 13, no. 2, pp. 269–282, 2020.

[11] Y. Lecun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[12] Y. Bengio, A. Courville, and P. Vincent, "Representation learning: a review and new perspectives," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 8, pp. 1798–1828, 2013.

[13] K. Fukushima, S. Miyake, and T. Ito, "Neocognitron: a neural network model for a mechanism of visual pattern recognition," *IEEE transactions on systems, man, and cybernetics*, vol. SMC-13, no. 5, pp. 826–834, 1983.

[14] Y. LeCun, B. Boser, J. S. Denker et al., "Backpropagation applied to handwritten zip code recognition," *Neural Computation*, vol. 1, no. 4, pp. 541–551, 1989.

[15] G. E. Hinton, "Learning multiple layers of representation," *Trends in Cognitive Sciences*, vol. 11, no. 10, pp. 428–434, 2007.

[16] G. E. Hinton and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *Science*, vol. 313, no. 5786, pp. 504–507, 2006.

[17] G. Dahl, M. Ranzato, A. Mohamed, and G. Hinton, "Phone recognition with the mean-covariance restricted Boltzmann machine[J]," *Advances in Neural Information Processing Systems*, vol. 23, pp. 469–477, 2010.

[18] L. Li Deng, "The MNIST database of handwritten digit images for machine learning research [best of the web]," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 141-142, 2012.

[19] T. N. Sainath, B. Kingsbury, G. Saon et al., "Deep convolutional neural networks for large-scale speech tasks," *Neural Networks*, vol. 64, pp. 39–48, 2015.

[20] Y. Qian, M. Bi, T. Tan et al., "Very deep convolutional neural networks for noise robust speech recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 12, pp. 2263–2276, 2016.

[21] G. Hinton, L. Deng, D. Yu et al., "Deep neural networks for acoustic modeling in speech recognition: the shared views of four research groups[J]," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 82–97, 2012.

[22] S. Sun, B. Zhang, L. Xie, and Y. Zhang, "An unsupervised deep domain adaptation approach for robust speech recognition [J]," *Neurocomputing*, vol. 257, pp. 79–87, 2017.

[23] Y. Freund, R. Iyer, R. Schapire, and Y. Singer, "An efficient boosting algorithm for combining preferences[J]," *Journal of Machine Learning Research*, vol. 4, no. Nov, pp. 933–969, 2003.