Hindawi
Mobile Information Systems
Volume 2022, Article ID 2262696, 11 pages
https://doi.org/10.1155/2022/2262696

Research Article

Application of Multimodal NLP Instruction Combined with Speech Recognition in Oral English Practice

Jin Xu and Tong Li

School of Foreign Languages and Literature, Tianjin University, 300350, China

Correspondence should be addressed to Tong Li; tongli@tju.edu.cn

Received 15 March 2022; Revised 4 April 2022; Accepted 11 April 2022; Published 27 April 2022

Academic Editor: Chia-Huei Wu

In order to study the application of multimodal NLP instruction combined with speech recognition based on hybrid deep learning in oral English practice, firstly, the basic principle of speech recognition technology is introduced. The concept of hidden Markov model and three key algorithms are explained, and its simulation and implementation in speech recognition application are realized. The architecture and key technologies of the system are introduced. Then, it introduces the specific application of deep learning in NLP. Finally, Chinese teachers with oral English teaching experience participate in the recording. The effective reading time of each person is 65 minutes, and the reading sentences are 3100 sentences. The total number of people is 80 (40 men and 40 women). The sentences cover 1595 spoken English words. Conduct oral English training. The experimental results show that the recognition accuracy decreases by about 2%, but the recognition speed increases by 10 times. In addition, the scoring accuracy is equivalent to that of the platform system. The accuracy of this method in instruction classification is increased, which verifies the feasibility and effectiveness of this method. In the future, attention mechanism will be used to expand this method.

1. Introduction

With the rapid development of international trade integration and China's trade opening, China has more and more exchanges with other countries in the world. Learning and mastering foreign languages, especially English, have become an important tool for human life and work. With the rapid growth of demand for English language learning, more and more language schools, teaching aids, and teaching materials have been launched one by one. However, oral language teaching has always been a difficult problem for Chinese people to learn English. The main reasons are the following two aspects: (1) there are great differences between the characteristics of Chinese pronunciation and English pronunciation, which makes Chinese people who learn a foreign language under the deep influence of their mother tongue make many pronunciation mistakes that are difficult or impossible to detect. (2) There is a shortage of foreign language teachers in China. Even primary and secondary schools in large and medium-sized cities do not have their own standardized language and English teachers who can teach English well. General information is only taught individually and not on a student-by-student basis. Both teachers and students can do oral teaching, so it cannot play a very effective role [1].

There is an extreme lack of qualified oral foreign language teachers in China. Even primary and secondary schools in large and medium-sized cities lack English teachers who have their own pronunciation standards and can accurately guide oral English learning. The general media teaching can only be taught unilaterally, but not according to the specific situation of students. Teachers and students can carry out oral teaching interactively, so it cannot play a very effective role [2].

At present, computer-aided language learning systems mostly focus on the learning of words and grammar. There is only some oral learning software with single function, which can only give learners an overall score of pronunciation. However, due to the limitation of their own level, it is difficult for self-scholars to find errors and correct incorrect pronunciation by themselves [3]. The use of speech recognition technology (Figure 1) equips the software with the ability to correct speech errors, which can help students learn to

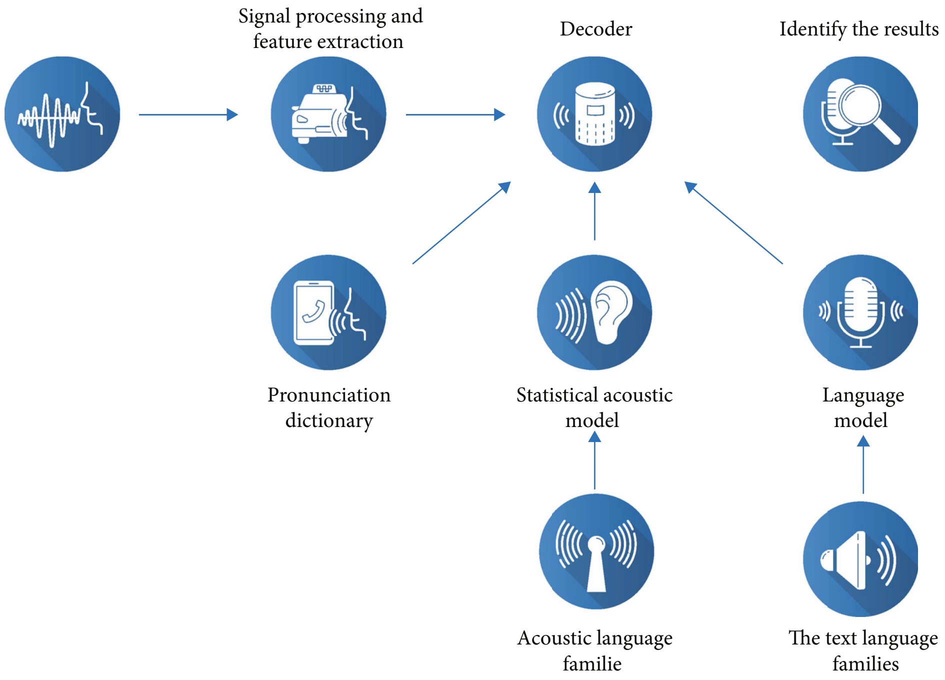
Signal processing and feature extraction
Decoder
Identify the results
Pronunciation dictionary
Statistical acoustic model
Language model
Acoustic language familie
The text language families

Figure 1: Speech recognition technology.

correct inaccuracies and avoid repeated mistakes. Greatly improving the efficiency of learners' oral learning will obtain great social benefits and market value. At present, some products developed by speech recognition technology have gradually begun to enter the current market, such as voice vehicle navigation system, intelligent robot, voice controlled interactive product echo, intelligent voice assistant on mobile phone, and voice input method. These applications have far-reaching significance for improving people's living standards. In short, as an important research direction of artificial intelligence, speech recognition technology is a research hotspot in the current society.

NLP is the abbreviation of Neurolinguistic Programming. N stands for nerve, L for language, and P for program. NLP studies the processes by which language affects the mind and body. In the 1970s, its founders Richard Bandler and John Grinder worked together on a topic: how people affect others, what are the common characteristics of efficient people, and how to replicate their behavior (Revell and Norman 1999).

Their main research objects are three masters with high attainments in the field of psychotherapy: family therapy masters Virginia Sutter and Gestalt founder Fritz Perls and hypnotherapy master Milton Erikson. Richard Bandler and John Grinder are committed to studying these outstanding people in psychology, studying their language and thinking mode, discovering their common ground, and reducing it to a set of programs that can be copied and imitated, which can be learned and imitated by willing people. It can be used not only for psychotherapy but also for self-improvement. The core of NLP is "imitation." There are four main principles in NLP (O'Connor and McDermott, 1996): first, goal setting (clarifying the needs of yourself and others); second, consistent affinity (establishing affinity with yourself and others); third, sensory acuity (mobilizing multiple senses to play a role); and fourth, the flexibility of behavior (flexible use of methods) [4]. The application of NLP in English teaching not only helps to improve students' learning efficiency and interest in learning but also helps students overcome the psychological obstacles of oral English expression.

2. Literature Review

Schuller found that at present, College English teaching generally pays attention to "reading and writing" ability and ignores the cultivation of "listening and speaking" ability. Language learning requires not only the ability to write and translate but also the ability to speak more often. There are many students who have good writing skills but cannot open their mouths or speak English. It has caused many college students to "mute" English [5]. According to Rodman et al., there is an extreme lack of excellent oral foreign language teachers in China. In China, even in large and medium-sized cities, few schools specially employ oral foreign teachers or English teachers with standard pronunciation, and there is a serious shortage in rural areas. The general media teaching, which can only be taught unilaterally and cannot be aimed at the specific situation of students, cannot play an effective role, because it cannot carry out oral teaching like the interaction between teachers and students and cannot provide accurate guidance [6]. Song's research suggests that there is a lack of a good atmosphere for practicing oral English after class. Domestic schools basically do not set up places like "English corner" for students to practice oral English. There is no learning atmosphere for students to practice oral English actively. Moreover, domestic English teaching pays more attention to English reading and word learning, and oral learning is not paid much attention. Oral examination is basically not set in primary, middle, and high schools [7]. Toledo et al.'s research suggests that NLP believes that behavior can change ideas. If you want to be confident, you must first be confident. Confident students speak loudly, clearly, and fluently. These advantages can be regarded as the common characteristics of students with excellent oral performance, and other students can learn and imitate these characteristics [8]. Calandruccio et al.'s research points out that only by clearly knowing what you want can you better work for it and be motivated. NLP is unique in that it advocates not only knowing what you want to achieve, but also knowing what others want. The combination of the two can achieve good results. Therefore, according to the guiding guidance of this concept, teachers teaching oral English should not only think about "what teaching objectives I want to achieve," but also consider "what learning objectives students want to achieve." The teaching effect achieved by this thinking mode is quite different from simply considering "what problems students have" [9]. The context aware selection method of multimodal dialogue filler in man-machine dialogue proposed by Nordström and Laukka adopts a Bayesian model to sample the filling time and collect the context information during the dialogue [10]. Yazdani proposed to enhance TED-LIM corpus through facial information, context text, and object recognition, which laid the foundation for robot multimodal natural language understanding method. However, most of the spoken language understanding (SLU) methods used in DSR are still rule-based [11].

3. Method

3.1. Application of Speech Recognition in Oral English Practice. Now, most of what we know about phonetics is the principle of comparison. According to this principle, the model of working presentation should be compared with the design of presentation skills one by one, and the best matching design should be adopted according to the results. The general recognition process is divided into presignal communication, eigenvalue extraction, training modeling, matching modeling (recognition), result determination, and recognition.

The process of human speech is caused by the contraction of the lungs, compressed air, and oscillations of sound caused by bronchi passing through the glottis and airways. There are three different kinds of stimulation in human speech, so it can produce three different types of sound, namely, voiced, unvoiced, and staccato. Although humans can make many endless sounds, words use fewer numbers to convey meaning. Generally speaking, a language has only a dozen phonemes. A phoneme is equivalent to a set of encoded characters in a communication system, consisting of a finite number of characters. According to the different states of speech and action, we can divide phonemes into open and closed. Closed phonemes are consonants in English pronunciation, and open phonemes are vowels. Some vowels, though simple in tone, have a narrower pitch, producing a slight fricative sound called a semivowel [12].

The speaker signal is an analog signal whose amplitude changes over time. After digitization, it can be recognized and processed by a computer. The digitalization of speech signal is the basis of digital processing. The process of digitizing speech signals includes testing and quantification. Through these two processes, digital signals of different amplitudes are obtained.

Because the signal is unstable, the process of speech signal is closely related to the movement of sound in the body. The body moves more slowly than sound vibrates. Therefore, speech signals can generally be considered to be very short in time; i.e., the spectral properties can remain roughly constant over periods of 10 to 20 milliseconds. The basic means of time-dependent processing is generally to intercept a speech signal with a limited length window sequence $\{w(m)\}$ for analysis [13] and let the window east to analyze the signal near any time. The general formula is as follows:

$$Qn = \sum_{m=-\infty}^{\infty} T[x(m)] * w(n-m), \tag{1}$$

where $T[]$ represents a certain operation and $\{x(m)\}$ is the input signal sequence. Equation (1) is in convolution form, so Q_n can be understood as the output of discrete signal $T[x(m)]$ through an FIR low-pass filter whose unit stimulus should be $\{w(m)\}$, as shown in Figure 2. Since the window function is generally taken as a smooth function with large middle and small ends of $x(n)$, the filter corresponding to such impulse response has low-pass characteristics. Its bandwidth and frequency response depend on the choice of creation function. The three most used window functions are rectangular window, Hamming window, and Hanning window, which are defined as

Rectangular window

$$w(n) = \left\{ \begin{matrix} 1 & 0 \le n \le L-1 \\ 0 & \text{other} \end{matrix} \right\}. \tag{2}$$

Hamming window

$$w(n) = \left\{ \begin{matrix} 1 & 0.54 - 0.46 \cos\left(\dfrac{2\pi n}{L-1}\right) \\ 0 & \text{other} \end{matrix} \right\}. \tag{3}$$

Hanning window

$$w(n) = \left\{ \begin{matrix} 1 & 0.5\left[1 - \cos\left(\dfrac{2\pi n}{L-1}\right)\right] \\ 0 & \text{other} \end{matrix} \right\}, \tag{4}$$

where L is the window length, and these window functions have low-pass characteristics. Comparative analysis shows that rectangular windows with high side lobes will cause large water leakage, so rectangular windows are rarely used, while Hamming windows with low side steps can overcome water leakage and have low characteristics, so they are most

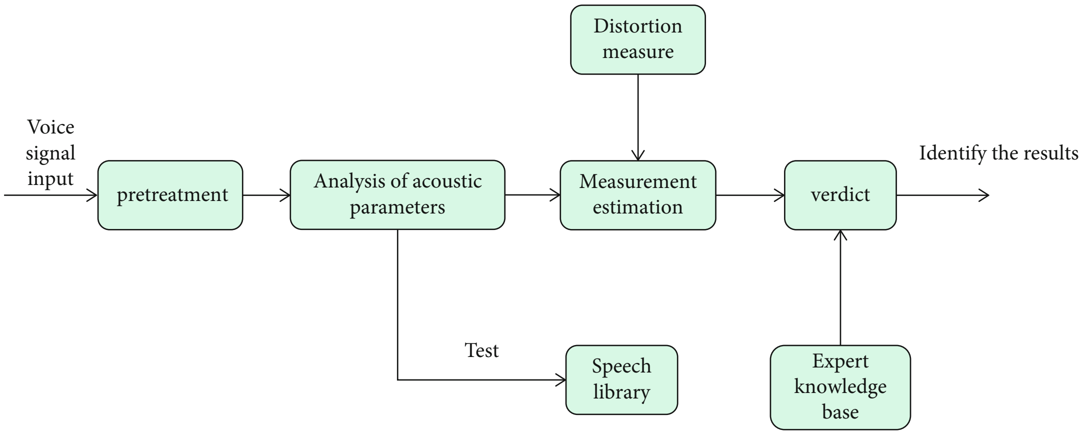

FIGURE 2: Flow chart of speech recognition.

used. In addition, the longer the window, the stronger the average interference to the signal, the higher the resolution of the signal [14], but the lower its resolution time. Therefore, in order to intercept different files at different speeds, the window length should be shorter (see Figure 3).

Voice endpoint detection detects the beginning and end of a speech. There are two common methods: multithreshold front-end endpoint detection method and double-threshold front-end endpoint detection method. In order to facilitate real-time removal, it is common to use both the initial discovery algorithm and the end discovery algorithm, because even multiple initial zeros are used on the front-end discovery value. The algorithm can reduce the forward error, and it has a long delay. This is not necessary for time management. However, using speechless time domain-short time no E (short time intensity) and Z (short time zero crossing value) for speech, the final search term can overcome many of the shortcomings of the starting finder [15].

Short term: the power of the speech signal changes visually over time. In general, the energy of voiceless speech is much less than that of speech and is therefore useful in exploiting the differences between voiceless and human voice, pitch, and voiceless parts. For signal $x(n)$, the short-time energy is defined as follows:

$$E_n = \sum_{m=-\infty}^{\infty} [X(n) * w(n-m)]^2 = \sum_{N=0}^{N-1} s_w^2(n). \qquad (5)$$

Because the short time of energy is the square function of the signal, there are differences between high and low signals, which is not suitable for integration in some applications. A simple way to solve this problem is to use a short-term average amplitude to express the change in energy, as shown in the following formula:

$$Mn = \sum_{m=-\infty}^{\infty} [X(n)w(n-m)]^2 = \sum_{m=n}^{n+N-1} |x_w(m)|. \qquad (6)$$

Short-term mean zero crossing: as the name implies, short-term mean zero crossing is the number of times the signal crosses zero in each post. For the difference, the short-term mean zero crossing value is as significant as the number of signal changes at the signal sampling point. It has two important applications: first, it is used to roughly describe the spectral characteristics of signals. The second is to judge the position of speech start and end point in combination with short-term energy, that is, endpoint detection, which is defined as the zero crossing rate of signal $\{x(n)\}$ which is defined as

$$Zn = \sum_{m=n}^{n+M-1} |\mathrm{sgn}\,[x(m) - \mathrm{sgn}\,[x(m-1)]]|\, w(n-m), \qquad (7)$$

where sgn [] is the symbolic function:

$$\mathrm{sgn}\,[x(n)] = \left\{ \begin{matrix} 1 & x(n) \geq 0 \\ -1 & x(n) < 0 \end{matrix} \right\}. \qquad (8)$$

And $w(n)$ is the window sequence, set to

$$w(n) \left\{ \begin{matrix} \dfrac{1}{2N} & 0 \leq N-1 \\ 0 & \text{other} \end{matrix} \right\}. \qquad (9)$$

The window amplitude here is $1/2N$, which means averaging the zero crossing numbers within the window range, because there are samples in the window, and each sample uses 2. Of course, you can also use other forms of windows instead of right-angle windows. In practical application, in order to avoid excessive zero crossing rate caused by random noise in mute section, a threshold is usually set first. When the symbols of the current and later two samples are different and the difference is greater than the threshold, the value of zero crossing rate is increased by 1 [16].

Mel frequency cepstral coefficient (MFCC) is one of the short-term acoustic characteristic parameters widely used in speaker recognition system. In theory, cepstrum parameters have obvious robustness. In addition, cepstrum parameters have two obvious advantages. One is that the spectrum can be processed by filtering and weighting the cepstrum domain. The second advantage is that Mel cepstrum theory

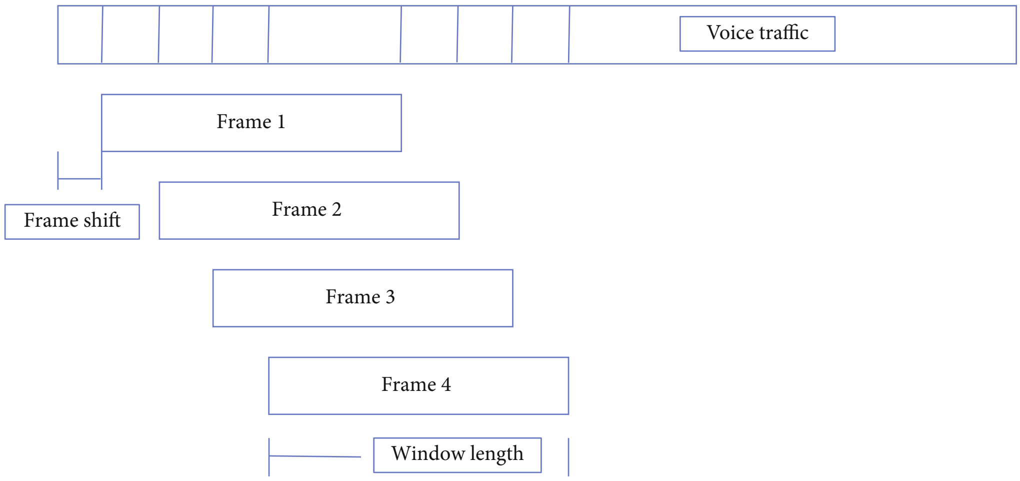

Figure 3: Intercepting voice frame with rectangular window.

can be easily applied. Different from ordinary cepstrum analysis, Mel frequency cepstrum parameter analysis focuses on the auditory mechanism of human ear and analyzes the spectrum of speech according to the results of auditory experiment, to obtain high recognition rate and good noise robustness [17].

One of the reasons why human ears can hear speech signals from noisy background noise is that human inner ear basement membrane can regulate external signals. For different frequencies, the signals in the corresponding critical bandwidth will cause vibration at different positions on the basement membrane. Thus, the band-pass filter bank can be used to simulate human ear hearing, to reduce the impact of noise on speech. First, let us explain the concept of critical frequency band. It is found that when the sound pressure is constant, when the noise is limited within a certain bandwidth, the subjective loudness felt by human ears is constant, and once the noise breaks through this bandwidth, the change of subjective loudness will be perceived. Similarly, when the sound pressure is constant, the loudness of a signal with complex envelope in this bandwidth is equivalent to the loudness of a pure tone at the central frequency of this bandwidth, which is independent of the frequency distribution of the signal itself. However, when the bandwidth of the signal breaks through the critical bandwidth [18], its loudness is no longer equivalent. According to Zwicker's work, the critical bandwidth changes with the change of frequency, which is consistent with the increase of perceived frequency. It is roughly linearly distributed below 1000 Hz, and the bandwidth increases logarithmically above 1000 Hz. Frequency describes the nonlinear relationship between human ear perception of frequency, and its relationship with frequency can be approximately expressed by the following formula:

$$\begin{aligned} \mathrm{Mel}(f) &= 1126.010471n\left(\frac{1+f}{700}\right), \\ \mathrm{Mel}^{-1}(f_{\mathrm{mel}}) &= 600\left(e^{ef_{\mathrm{mel}}/1126.01047} - 1\right). \end{aligned} \tag{10}$$

It can be seen from Figure 4 that template matching methods are often used similarly to count patterns in various systems of knowledge. In the training process, after feature extraction and feature dimension compression, clustering or other methods are used to generate one or more templates for each model class. At the acceptance level, feature vector similarity of the model should be recognized, each model should be calculated, and then, it is determined which class it belongs to. Speech recognition can also use comparison models to measure similarity, but there is an assembly time problem on a particular dimension, which is somewhat unusual and can be used for standard assurance comparisons. In this section, we will focus on a strictly stochastic signal model known as the hidden Markov model (HMM). We will first introduce Markov chain theory and then use several simple examples to expand the understanding of hidden Markov model. Then, we will focus on three basic problems of hidden Markov model design [19].

Speech signal is a quasistationary signal. HMM is a statistical model that can not only describe the dynamic changes of speech signal features but also well describe the statistical distribution of speech features. It is an excellent tool for quasi-static time-varying speech signal analysis and speaker recognition. This is an example that emerges from absence to describe the characteristics of random processes. It evolved from a chain. In phonetic knowledge, both random phonetic knowledge and phonetic knowledge are limited. It can be a one-dimensional sequence of observing or encoding characters or a multidimensional vector sequence. For example, a language fragment such as a word, phoneme, or phrase can be represented by a string of vectors, which is a diagnostic vector. If a string of vectors is a quantized vector, then each vector is represented by a coded symbol [20], which turns out to be a sequential analysis of symbols. Whether it is an observation vector sequence or an observation symbol sequence, it is collectively referred to as the observation sequence, which is recorded as

$$O = o_1 o_2 \cdots o_T. \tag{11}$$

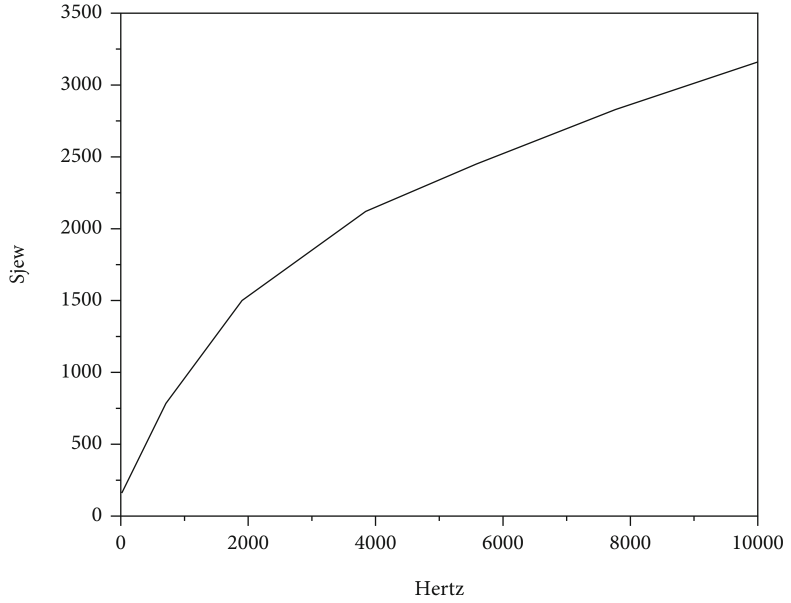

FIGURE 4: Relationship between Mel and frequency.

It can be seen from Tables 1 and 2 that among the above parameters, π and A determine the different structures of hidden Markov chains in HMM. In the left to right model with crossing, the transition state can only jump from left to right, not vice versa, as shown in Figure 5(a). This model has a small amount of computation and is very suitable for modeling speech signals, because the properties of speech signals change with time. Figure 5(b) is a more common and simpler Markov chain. It has no state crossing, so it becomes a left to right model without crossing.

3.2. Voice Scoring and Error Correction. The research content of oral English learning system based on speech recognition is very extensive. Some focus on the common pronunciation errors of beginners, such as various similar vowels and nasal sounds. Some focus on the unique pronunciation skills or difficulties of English pronunciation, such as intonation, continuous reading, and stress. Another kind focuses on the whole of an oral English learning system, which humanizes and optimizes the performance of the system according to the phonetic teaching method and computer. Speech recognition is the key to pronunciation learning, but it can directly improve English pronunciation. On this basis, many new studies need to be done:

(1) Look for ways to evaluate pronunciation. It is not only applicable to phoneme units but has also quantitative measurement standards for the performance indicators of hyperphoneme pronunciation vectors such as rhythm. The difficulty of this problem lies in how to calculate the pitch, stress, speed, and rhythm, as well as the relationship between their corresponding vocal acoustic structural features. Rate the pronunciation of words, phrases, or sentences [21]

TABLE 1: HMM parameters.

Model parameter	Explain
N	Number of states of the model
$A = \{a_{ij}\}$	State transition matrix
$\pi = \{\pi_i\}$	The initial probability distribution for each state
$B = \{b_j(o)\}$	Output the probability density function

(2) How to detect and correct phoneme level pronunciation errors of a given voice and give learners correction feedback in a friendly way

The level of learners' pronunciation is an important information fed back to learners by a computer-aided pronunciation learning system. Therefore, automatic English pronunciation scoring is the core and basic function of this kind of learning system. There must be a reference or standard for measurement. The two commonly used methods are based on the HMM trained by reference speech and reference corpus.

Eloquence evaluation not only involves telephone, description, environment, and other disciplines but also has emotional, physical, and cultural effects. This is a very difficult problem. According to the key score, the good speech test is divided into scores and target scores. Now, there are many ways to score. The main test points include intermediate test score, test combination decision, intermediate test score distortion, and test satisfaction decision. The main score for a speech is not only time and effort but also the accuracy of various factors. The test conditions and content of the examinee will affect the reliability of the test results for specific conditions. Therefore, in general, the use of tools to measure speech quality not only is affected

Table 2: HMM identification process parameters.

Parameter	Explain
O	Observation vector
M	The number of Gaussian elements contained in each state
c_j	The weight of the first mixed Gaussian in the J state
N	Represents the normal Gaussian probability density function
μ_{ij}	The mean vector of the first mixed Gaussian element in the J state
U_{ij}	The covariance matrix of the first mixed Gaussian element in the J state

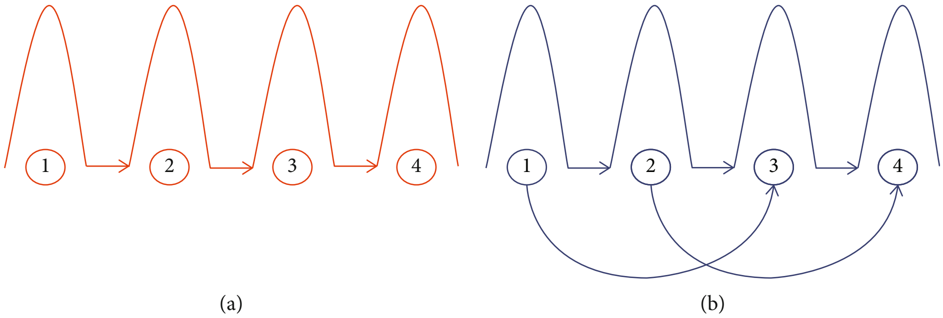

Figure 5: HMM state diagram.

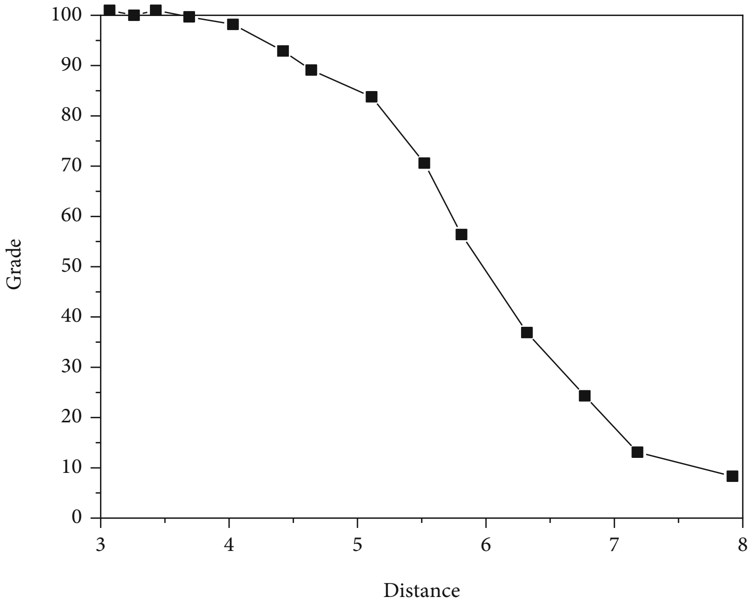

Figure 6: Calculated distance score conversion diagram.

by the environment and human context but also changes in many aspects of application. The time difference and the results of the time difference are also easy to compare directly. Now, there are many ways to measure good speech. Common ones are scores based on dynamic time warpage (DTW), scores based on HMM log purchases, scores based on HMM log subsequent results, scores based on segmented distribution, scores based on long term, scores based on performance, score probabilities based on reliability time, and so on. The above measures use the speech model as a model for various similar calculations [22].

As can be seen from Figure 6, an optimized algorithm is adopted to align the features with the template features by unevenly distorting and bending the time axis of the speech signal to be recognized and continuously calculate the matching path with the smallest distance between the two vectors, to obtain the regularization function with the smallest cumulative distance when the two vectors are matched,

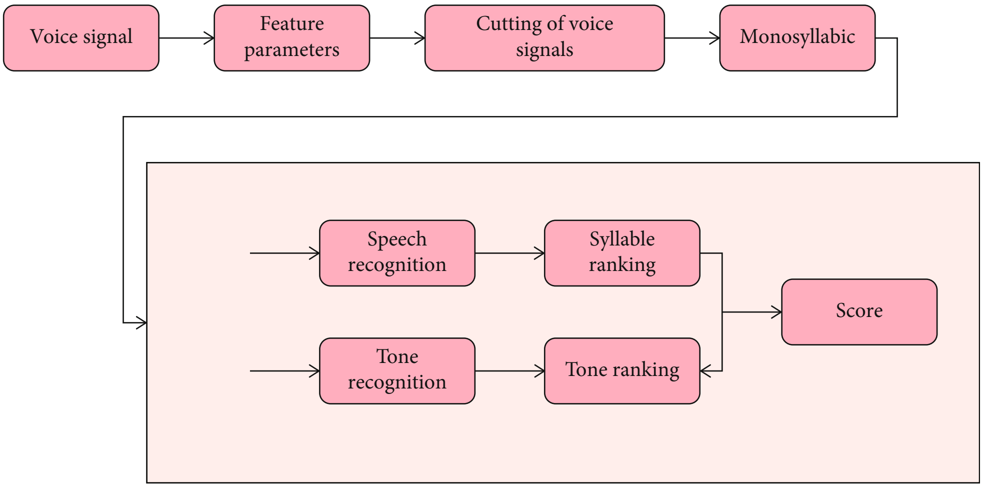

Figure 7: Flow chart of the scoring system.

which is the earliest and most used method to successfully solve the problem of speech pattern matching [23].

The scoring of HMM Speech model is another speech scoring method, which mainly starts from the two aspects of sound and tone, hoping to find out the difference between the test speech and the acoustic model and tone model and score the speech according to the difference.

The flow of the scoring system is shown in Figure 7. Taking the pretrained acoustic model and tone model as the standard answer, using speech recognition technology, find out the difference between the test speech and the model, and give the score with the scoring mechanism. In terms of feature parameter extraction, it mainly involves two feature parameters: fundamental frequency trajectory and Mel cepstrum parameters, which are used as the feature parameters of tone recognition and sound recognition, respectively. In the actual operation, Viterbi decoding is used to cut the speech signal into monosyllabic segments, and then, the sound model and tone model are compared for each syllable, and the recognition results are combined with our predesigned scoring mechanism to convert the score, that is, the score of the comparative test speech. This scoring system includes the commonly used technologies of speech recognition, such as the hidden Markov model (HMM), Tree Net, and Viterbi Algorithm. In terms of tone, it includes such aspects as orthogonal expansion, Chebyshev approximation, K-means clustering method, and classifier design.

3.3. Specific Application of Deep Learning in NLP. The core of NLP is imitation (Revell and Norman 1999). Imitation is the shortcut to success. How to get useful resources from others depends on imitation. Children's babbling is imitation. They learn to speak by imitation. The improvement of English learners' oral expression ability also depends on imitation. Just as every successful person has some similar characteristics that can be imitated by the general public, students with excellent oral English also have all aspects imitated by other students. Imitate the pronunciation and intonation of the target language and the thinking and behavior patterns of the target language population. It is necessary and effective to imitate pronunciation and intonation. First, because English and Chinese belong to different language families, their pronunciation also has different characteristics. Some syllables in English and Chinese look the same, and the actual pronunciation methods and parts are slightly different, such as consonants /f/ and /l/. In the Chinese pronunciation system, /f/ is a clear fricative sound of lips and teeth, with the lower lip close to the upper incisors to form a gap. The soft palate rises, and the nasal passage is closed. Make the air flow rub through the gap formed by the tooth lip to make a sound. In the English pronunciation system, /f/ is the friction consonant between the lips and teeth. When pronouncing, the lower lip gently touches the upper teeth, and the air flow passes between the lips and teeth to form a friction sound. There are also some sounds in the English pronunciation system that are not found in the Chinese phonology system, such as stops /θ/. When learning these phonetic symbols, we should pay more attention to practice in oral English practice. Practice this pronunciation method different from Chinese, so that the muscles of students' pronunciation organs can adapt to the English pronunciation system.

3.3.1. Application Process Description. The application of deep learning in natural language processing requires the scientific application of gradient descent method. The actual application process is as follows: (1) establish the corresponding model framework. Combined with the relevant contents that should be processed, ensure the rationality of the selected neural network structure, and achieve the purpose of establishing the corresponding deep learning model framework. (2) Check the model carefully. Reasonably use the gradient descent method to complete the task of checking the model, check and analyze the existing loopholes, and make sure whether they meet the relevant provisions. (3) Realize the initialization effect of the model. After careful inspection, scientifically optimize the relevant models in order to make up for the loopholes and defects, and scientifically improve the parameters of the relevant models. (4) Continuously improve relevant models. The regularization

Table 3: Data acquisition specifications.

Sampling rate	Track	Quantification length	Bit rate	Coded format	Collecting device	Sensor type
16K	Single	16 bit/s	256 bit/s	PCM	MAYA	MKH 8020

method is used reasonably, and the model parameters that do not meet the relevant provisions are improved in time, to meet the relevant fitting provisions.

3.3.2. Analysis of Application Measures

(1) Do a good job in marking word segmentation and part of speech

For word segmentation, according to relevant regulations, it can achieve the effect of recombining to continue the word order and combine it into a new word sequence at the same time. When tagging part of speech, ensure the accuracy of part of speech tagging. For example, this word is an adjective, verb, etc. By strengthening the application of deep learning method, we can carry out part of speech tagging, semantic role tagging, named entity identification, and so on.

(2) Scientific parsing syntax

That is, reasonably analyze the grammar of sentences and the relationship between different grammars. The scientific application of deep learning method can achieve the purpose of identifying sentence syntactic units in an automatic form, sort out the connections between different syntactic units, input a given sentence scientifically, make rational use of the characteristics of grammar, complete the task of establishing phrase structure tree, and deal with it effectively.

(3) Study the meaning of words carefully

In the process of using deep learning, we need to pay attention to the learning of word meaning and play a good role of relevant unsupervised learning system. In the process of establishing the deep neural network model, we should use this model scientifically and analyze it scientifically with reference to the context in the text, to obtain the best expression form of word meaning, master the vocabulary implied by word meaning, and achieve the purpose of accurately analyzing ambiguous words with the same name. If there are multiple polysemy word vectors, the method of model optimization can be adopted to enrich the semantics of word vectors and ensure the accuracy of expression.

(4) Strengthen the scientific analysis of emotion

In the process of reasonably using the deep learning method to analyze emotion, it is necessary to establish the corresponding emotion analysis model and, with the help of the training part of the deep neural network, effectively complete the task of labeling relevant emotion labeled sentences and refer to the corresponding laws and context characteristics, to achieve the effect of predicting the emotional characteristics of the marked sentences. Then, further analyze the emotional color of document level and sentence level. Obviously, this measure can play a good role in advanced affective analysis and improve the overall efficiency of natural language processing by effectively using deep learning methods.

4. Experimental Analyses

From the perspective of training template, corpus is the fundamental source of speech knowledge required by speech recognition engine. From the perspective of performance evaluation, the quality of the corpus will directly affect the scientificity and effectiveness of the evaluation results. A standardized and comprehensive corpus should do the following [23, 24]:

(1) Universality: the content is extensive, covering various voice phenomena as much as possible

(2) Representativeness: the speaker is widely representative in terms of gender, age, region, speaking speed, etc.

(3) Consistency: the corpus is marked in detail and consistent with the pronunciation content

Because the system does research on Chinese English pronunciation, the personnel participating in the corpus recording of the system are Chinese teachers with professional oral English teaching experience. The effective reading time of each person is 65 minutes, and the reading sentences are 3100 sentences. The total number of people is 80 (40 men and 40 women). The sentences cover 1595 spoken English words. A specially assigned person shall mark the time of each sentence at the word level. See the following table for specific recording equipment and data information.

Next, it introduces the characteristics of different learning strategies (see Table 3 for details).

(1) Free choice of class hours: under this strategy, the learning content will be fully displayed in front of users regardless of class hours

(2) 30-day plan: under this strategy, the learning content will be divided into 30 class hours, with 30 sentences in each class hour, which is helpful for users to learn in stages

(3) Intensive review: under this strategy, fill-in-the-blank multiple-choice questions of some sentences in the learning content will appear, so that users can recognize the answers by voice

Table 4: Continuous HMM correctness test cases.

	Data information	
Data type	Close testing	Open testing
Number of test cases	8000	5400
Number of speakers	11	6
Speaker ratio of male to female	2 : 1	3 : 2
Test data format	16 KHz/mono/16 bit/PCM	

Table 5: Correctness test results of continuous HMM.

	Test result	
	Close testing	Open testing
Male speaker	96.43%	94.13%
Female speaker	94.81%	92.89%
Gather	95.73%	95.54%

Table 6: Semicontinuous HMM correctness test cases.

	Data information	
Data type	Close testing	Open testing
Number of test cases	22500	8200
Number of speakers	26	10
Speaker ratio of male to female	1 : 1	4 : 5
Test data format	16 KHz/mono/16 bit/PCM	

Table 7: Correctness test results of semicontinuous HMM.

	Test result	
	Close testing	Open testing
Male speaker	95.72%	93.12%
Female speaker	91.46%	94.02%
Gather	93.68%	93.43%

(4) Intelligence enhancement: under this strategy, the software will conduct intensive training for previous error-prone problems.

HMM is the core of the system. It tests the recognition rate and recognition time performance of HMM and compares the performance of continuous HMM used in PC platform system and semicontinuous HMM used in this system. The experimental results are shown in Tables 4–7.

This test tested the function of the English language design model and successfully completed the operation of each module. By analyzing system terminology and comparisons and by evaluating performance in real time, the performance of the system shows that the system is efficient and able to meet customer needs. The results using the data show that speech recognition has been tested and evaluated, and the test results are satisfactory.

5. Conclusion

The application of NLP in college oral English teaching enriches the existing oral English teaching methods and widens the space for teachers to choose teaching methods in classroom design. The four main principles of NLP are conducive to helping teachers set teaching objectives more effectively, build an affinity classroom atmosphere, consciously mobilize students' sensory organs, and carry out flexible and interesting teaching activities; The relevant concepts of NLP are conducive to improving students' self-confidence in learning oral English, arousing students' awareness of self-acceptance, helping students clarify their learning objectives, and improving students' enthusiasm and initiative to participate in the classroom. While NLP improves students' oral ability, the further enrichment and development of its concept also help to eliminate the phenomenon of "dumb English" and help oral English teaching out of the current dilemma. Spoken English learning is an Android app developed by native speakers to meet the needs of users to properly learn and practice English anytime, anywhere. The system is developed by selecting some necessary tasks that need to be completed in the terminal, providing users with a simple structure for English learning and practice, including proficiency in speech, speech measurement, radio broadcasting, and oral communication.

In the use of spoken English, speech recognition is greatly influenced by the environment, and the influence of environmental noise reduces the recognition degree of the system. Although this paper uses the final recognition process to eliminate part of the Gaussian white noise, it still does not eliminate the noise and affect the speech skills. He also needs to learn and use good speech techniques. In recent years, products based on phonetic knowledge have been implemented in many fields. More and more scientists are investing in the science of speech therapy. Looking into the future, speech ability will be greatly improved, making people's lives easier and promoting the progress of knowledge.

Data Availability

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

Conflicts of Interest

The authors declare no competing interests.

Acknowledgments

This work is supported by the Philosophy and Social Science Project in Tianjin "The Research on Artificial Intelligence in American Feminist Sci-Fi Literatures" (No. TJWW21-006).

References

[1] K. A. Berg, J. H. Noble, B. M. Dawant, R. T. Dwyer, and H. Gifford, "Speech recognition as a function of the number of channels for an array with large inter-electrode distances," *The*

Journal of the Acoustical Society of America, vol. 149, no. 4, pp. 2752–2763, 2021.

[2] L. Calandruccio, P. A. Wasiuk, E. Buss, L. J. Leibold, and J. Oleson, "The effect of target/masker fundamental frequency contour similarity on masked-speech recognition," *The Journal of the Acoustical Society of America*, vol. 146, no. 2, pp. 1065–1076, 2019.

[3] R. Yazdani, J. M. Arnau, and A. Gonzalez, "A low-power, high-performance speech recognition accelerator," *IEEE Transactions on Computers*, vol. 68, no. 12, pp. 1817–1831, 2019.

[4] T. Hodgson, F. Magrabi, and E. Coiera, "Evaluating the usability of speech recognition to create clinical documentation using a commercial electronic health record," *International Journal of Medical Informatics*, vol. 113, pp. 38–42, 2018.

[5] B. W. Schuller, "Speech emotion recognition," *Communications of the ACM*, vol. 61, no. 5, pp. 90–99, 2018.

[6] C. Rodman, A. C. Moberly, E. Janse, D. Bakent, and T. N. Tamati, "The impact of speaking style on speech recognition in quiet and multi-talker babble in adult cochlear implant users," *The Journal of the Acoustical Society of America*, vol. 147, no. 1, pp. 101–107, 2020.

[7] Z. Song, "English speech recognition based on deep learning with multiple features," *Computing*, vol. 102, no. 3, pp. 663–682, 2020.

[8] T. D. Toledo, H. D. Lee, N. Spolaor, C. R. Coy, and F. C. Wu, "Web System Prototype based on speech recognition to construct medical reports in Brazilian Portuguese," *International Journal of Medical Informatics*, vol. 121, pp. 39–52, 2019.

[9] L. Calandruccio, E. Buss, P. Bencheck, and B. Jett, "Does the semantic content or syntactic regularity of masker speech affect speech-on-speech recognition?," *The Journal of the Acoustical Society of America*, vol. 144, no. 6, pp. 3289–3302, 2018.

[10] H. Nordström and P. Laukka, "The time course of emotion recognition in speech and music," *The Journal of the Acoustical Society of America*, vol. 145, no. 5, pp. 3058–3074, 2019.

[11] R. Yazdani, J. M. Arnau, and A. Gonzalez, "Laws: locality-aware scheme for automatic speech recognition," *IEEE Transactions on Computers*, vol. 69, no. 8, pp. 1197–1208, 2020.

[12] X. Cui, W. Zhang, U. Finkler, G. Saon, M. Picheny, and D. Kung, "Distributed training of deep neural network acoustic models for automatic speech recognition: a comparison of current training strategies," *IEEE Signal Processing Magazine*, vol. 37, no. 3, pp. 39–49, 2020.

[13] S. Hu, X. Shang, Z. Qin, M. Li, C. Wang, and C. Wang, "Adversarial examples for automatic speech recognition: attacks and countermeasures," *IEEE Communications Magazine*, vol. 57, no. 10, pp. 120–126, 2019.

[14] J. Han, Z. Zhang, G. Keren, and B. Schuller, "Emotion recognition in speech with latent discriminative representations learning," *Acta Acustica united with Acustica*, vol. 104, no. 5, pp. 737–740, 2018.

[15] M. Maslowski, A. S. Meyer, and H. R. Bosker, "Listeners normalize speech for contextual speech rate even without an explicit recognition task," *The Journal of the Acoustical Society of America*, vol. 146, no. 1, pp. 179–188, 2019.

[16] J. Zhao, M. Xia, and L. Chen, "Learning deep features to recognise speech emotion using merged deep cnn," *IET Signal Processing*, vol. 12, no. 6, pp. 713–721, 2018.

[17] B. Widanski, J. A. Thompson, and K. Foran-Mulcahy, "Improving students' oral scientific communication skills through targeted instruction in organic chemistry lab," *Journal of Chemical Education*, vol. 97, no. 10, pp. 3603–3608, 2020.

[18] A. Galve, "Zygoma quad compared with 2 zygomatic implants – a systematic review and meta-analysis," *Clinical Oral Implants Research*, vol. 30, no. S19, pp. 402–402, 2019.

[19] H. Peng, Y. Ma, Y. Li, and E. Cambria, "Learning multi-grained aspect target sequence for Chinese sentiment analysis," *Knowledge-Based Systems*, vol. 148, pp. 167–176, 2018.

[20] J. Abhyuday, L. Feifan, L. Weisong, and Y. Hong, "Overview of the first natural language processing challenge for extracting medication, indication, and adverse drug events from electronic health record notes (MADE 1.0)," *Drug Safety*, vol. 42, no. 1, pp. 99–111, 2019.

[21] O. Golani, D. Pilori, F. Guiomar, G. Bosco, A. Carena, and M. Shtaif, "Correlated nonlinear phase-noise in multi-subcarrier systems: modeling and mitigation," *Journal of Lightwave Technology*, vol. 38, no. 6, pp. 1148–1156, 2020.

[22] N. Majumder, S. Poria, H. Peng et al., "Sentiment and sarcasm classification with multitask learning," *IEEE Intelligent Systems*, vol. 34, no. 3, pp. 38–43, 2019.

[23] A. Y. Bonino and A. R. Malley, "Measuring open-set, word recognition in school-aged children: corpus of monosyllabic target words and speech maskers," *The Journal of the Acoustical Society of America*, vol. 146, no. 4, pp. EL393–EL398, 2019.

[24] A. Frimane, T. Soubdhan, J. M. Bright, and M. Aggour, "Nonparametric bayesian-based recognition of solar irradiance conditions: application to the generation of high temporal resolution synthetic solar irradiance data," *Solar Energy*, vol. 182, pp. 462–479, 2019.