Eng Int Syst (2019) 3: 141–150

Engineering
Intelligent Systems

Optional English Speech Teaching Method Based on Recognition Emotion Mining and Deep Learning Algorithms

Xinyu Zhang, Hui Li*, Na Wang and Ruolin Shi

Cangzhou Normal University, Cangzhou, Hebei 061000, China

Speech recognition technology and speech evaluation technology is the core of computer-aided speech learning. Of the two, speech recognition technology is particularly critical and plays a vital role. This paper analyses the application of emotion mining and deep learning algorithms in the recognition and teaching of English speech. Emotional visualization can enable teachers and curriculum managers to be more intuitive when sensing emotional changes in students' learning process, and will assist teachers to provide personalized teaching and intervention. The experimental results show that the model presented in this paper for English phonetics teaching, speech recognition and evaluation is reasonable and valid. It can give learners timely, accurate and objective evaluation and feedback guidance, and can help learners identify the differences between their own pronunciation and the standard pronunciation. In addition, this paper optimizes the English language model by means of the sub-word modeling method, which alleviates the problem of sparseness and robustness of the traditional whole-word language model brought by the very large vocabulary of adhesive words.

Keywords: Speech recognition; Deep learning; Emotion mining; DNN algorithm

1. INTRODUCTION

The core of Computer Aided Language Learning (CALL) is speech recognition technology and speech evaluation technology, with the former being the key. Speech recognition technology, namely Automatic Speech Recognition (ASR), is the technology whereby voice signals are converted into corresponding commands or texts through automatic recognition and understanding by machines (generally referred to as computers), in order to produce intelligent voice interaction between human beings and machines (Mushtaq et al., 2017; Wang et al., 2018). Therefore, speech recognition has become a popular research field in recent years. Because of the complexity of speech changes, the large number of speech signals, the many speech feature parameters, and the large amount of computation of speech recognition and evaluation, this requires a large amount of voice and signal processing, necessitating better software and hardware resources and algorithms (Bleisch et al., 2014). From the classic dynamic time warping algorithm (DTW) to the mainstream hidden Markov model (HMM) to the traditional artificial neural network, speech recognition technology has made much progress, but it has also encountered an unprecedented bottleneck (Zhang et al., 2016). It is difficult to make further improvements to its accuracy and speed; hence, the difficulty in making a breakthrough in the commercialization of speech recognition (Mishra et al., 2016; Upadhyay et al., 2018).

Chinese students generally study Chinese for several years and they have a certain Chinese foundation before they begin to learn English (Meltzner et al., 2017; Jesse et al., 2017).

*Corresponding author: Hui Li, Email:zxyminnie@163.com.

Therefore, English pronunciation is influenced by dialect and Mandarin. Students who have correct pronunciation and fluency when they first learn English are rare. In addition, there are some similarities between English and Chinese pronunciations (Stoean et al., 2010; Kang et al., 2015; Dorothy et al., 2017). Therefore, the standard of the student's Mandarin may directly affect the accuracy of his/her English pronunciation.

All languages in the world have three core elements: vocabulary, grammar, and speech. As a tool for conversation, language is important, of which pronunciation is the most important. (Cao et al., 2016). The purpose of learning English is to use this language as a tool for conversation, to accurately express one's thoughts and to understand others' thoughts (Liang et al., 2014; Mernik et al.,2015). Standard speech is fundamental to spoken expression and listening comprehension. Only English with a good voice system can learn English well.

The smallest and most natural unit of speech is the phoneme. Acoustically, the phoneme is divided in terms of sound quality (Mahboubi, 2014; Sun et al., 2015). Physiologically, a phoneme is formed by the action of pronunciation. For example: /eg/ contains two pronunciations of /e/ and /g/, which are two phonemes. The same phoneme is the sound formed by the same pronunciation action, and the different phonemes are the sounds formed by different pronunciation actions. For example, the two /d/ pronunciations in /hænd/ and /bænd/ are the same, so they are the same phonemes; the /b/ and /z/ are pronounced differently, so they are different phonemes. The analysis of phonemes is generally based on pronunciation actions. For example, the pronunciation action of /m/ is: the upper lip and the lower lip are closed, the vocal cords vibrate, and the air rushes out of the nasal cavity, which is classified as a lip nasal pronunciation (Wang et al., 2016).

The symbol used to record phonemes in phonetics is called the phonetic symbol. For example, international phonetic symbols are generally marked with "//". To clarify: "A phoneme is represented by only one phonetic symbol, and a phonetic symbol represents only one phoneme. The International Phonetic Alphabet (IPA) is developed by the International Phonetic Association and strictly stipulates the principle of 'one note and one tone', that is, 'one phoneme corresponds to one symbol and one symbol corresponds to one phoneme'. It is based on the Latin alphabet and is supplemented by a method of changing glyphs and borrowing letters from other languages". To ensure consistency, most symbols guiding the pronunciation still retain the original sound of Latin or other languages.

It is generally believed by phoneticists that speech can be divided into two categories: vowels and consonants, depending on whether the airflow is blocked by the vocal organs when exhaled from the lungs. When speaking, the unobstructed phoneme of the airflow is a vowel, and the phoneme whose airflow is obstructed is a consonant. The difference between a vowel and a consonant is that the airflow is unobstructed in the channel when the vowel is emitted, and the airflow in the channel is hindered to varying degrees when the consonant is emitted, that is, the airflow is closed or the airflow is narrowed. In addition, we can distinguish vowels and consonants in other ways: 1) Airflow. When the vowel is pronounced, the airflow is weaker. In contrast, the airflow is stronger when the consonant is emitted, especially when a voiceless consonant is emitted. 2) The tension of the articulated organ. When vowels are pronounced, the vocal organ tension is average. However, when a consonant is produced, the tension of the vocal organ occurs at the moment of pronunciation and occurs at a certain point in the utterance, while the other vocal organs do not exhibit a state of tension. 3) Sound. From the perspective of acoustic phonetics, vowels have greater loudness and intensity than consonants. 4) Music and noise.

Hence, it can be seen that the current English speech recognition technology is still in the research stage, and most of the research has certain deficiencies in terms of the fault tolerance rate. Based on this, this study constructed an English speech recognition system based on machine learning, and tested and evaluated the performance of the system.

2. RESEARCH METHOD

2.1 Hidden Markov (HMM) Acoustic Modeling

The conversion formula for the Mel domain frequency and the ordinary linear frequency is:

$$f_{mel} = 2959 \log_{10}\left(1 + \frac{f}{700}\right) \tag{1}$$

As shown in Figure 1, a typical HMM can be mathematically described using five sets of parameters, namely:

$$M = \{O, \Omega, \pi, \mathrm{A}, \mathrm{B}\} \tag{2}$$

In equation (2), O represents the sequence of observation vectors $\{o_1, o_2, \cdots, o_T\}$ from time 1 to time T, and Ω_0 represents the set $\{s_1, s_2, \cdots, s_K\}$ of finite implicit state sequences contained in the K HMMs. The model parameters that need to be determined in the HMM are represented by three sets of parameters: $\lambda = \{\pi \mathrm{A}, \mathrm{B}\}$. $\pi = \{\pi_1, \pi_2, \cdots, \pi_K\}$ denotes the distribution of K states in which the HMM is at the starting time, $A = \{a_{ij}\}_{K\times K}$ denotes the state transition probability matrix, and B denotes the probability distribution function $\{b_i(o)\}$ under different states. In speech recognition systems, the state output density function is usually characterized by a Gaussian Mixture Model (GMM), that is:

$$b_i(o_t) = \sum_{k=1}^{k} w_{ik} \cdot \frac{1}{\sqrt{(2\pi)^D |\sum_{ik}|}} \exp\left[-\frac{1}{2}(o_t - \mu_{ik})\right] \tag{3}$$

In the figure, a_{ij} represents the transition probability between states jumping from state i to state j. $b_i(o_t)$ represents the probability of outputting the observation vector o_t when jumping from state i.

The parameters of the HMM must satisfy the following conditions:

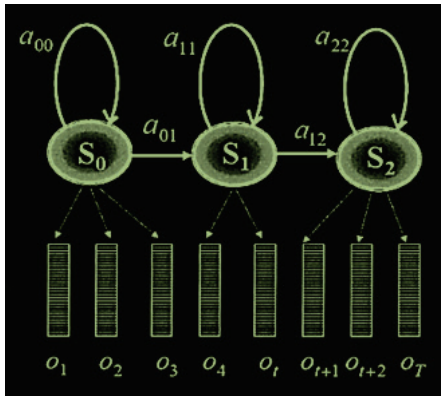

Figure 1 Schematic diagram of the first-order HMM acoustic model.

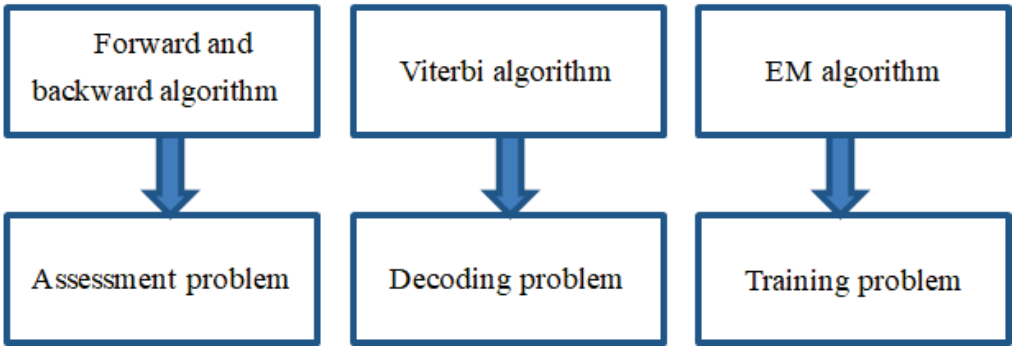

Figure 2 Corresponding image of the problem and the solution algorithm.

$$\pi_i \geq 0, a_{ij} \geq 0, b_i\left(o_t\right) \geq 0$$
$$\sum_{i=1}^{N} \pi_i = 1 \sum_{j=1}^{N} a_{ij} = 1 \int b_i\left(o_t\right) do_t = 1 \tag{4}$$

Considering its practical application, HMM should make two important assumptions: One is the first-order Markov hypothesis, that is, the state S_t of the current time t is only related to the state S_{t-1} at the previous moment $t-1$ and has nothing to do with any state at any other time. The formula is:

$$p\left(s_t|s_1^{t-1}\right) = p\left(s_t|s_{t-1}\right) \tag{5}$$

The other is the output-independent hypothesis, which means that the output value at the current time is only governed by the probability density of the current state and is independent of other output values and states that have already been generated. The formula is:

$$p\left(x_t|x_1^{t-1}, x_1^t\right) = p\left(x_t|s_t\right) \tag{6}$$

In order for HMM to better serve the speech recognition system, there are three classic problems that we need to solve:

(1) Probability calculation problem: In the case of given model parameters, how the probability $p\left(O|\lambda_{mm}\right)$ of the observed vector O appears is calculated.

(2) Code problem: Under the premise of the given observation vector and model parameters, how optimal the state sequence Q is calculated, so that the joint probability $p\left(O, Q|\lambda_{mm}\right)$ is the largest.

(3) Estimated problem of model parameter λ_{mm}: After giving enough observation vector sequences, how can we estimate the model parameters λ_{mm} using the existing data so that the estimated hidden Markov model is the most probable model for generating a given observation vector.

According to the HMM, any state sequence S may generate the observation vector O with a certain probability, so $p(O|\lambda_{mm})$ should be the cumulative sum of the probabilities corresponding to each possible state sequence, that is:

$$p(O|\lambda) = \sum_s p(O, s|\lambda) = \sum_s p(s|\lambda) p(O, s|\lambda) \tag{7}$$

Then, for an HMM with a total of M states, if the length of time is N, the number of all possible state sequences is M^N. Moreover, the computational complexity increases exponentially as the number of states and the length of time increase, which is unacceptable. The algorithm that effectively solves this complexity problem is the Forward-Backward Algorithm, which will be discussed below.

The definition of forward probability is:

$$\alpha_i(i) = p\left(o_1, o_2, \cdots, o_t, s_t|\lambda\right) \tag{8}$$

This means that the output vector at time 1 to t is $o_1, o_2, \cdots, o_t$ and the state at time t is the probability of i.

First, it is initialized:

$$\alpha_1(1) = 1 \alpha_1(j) = 0 (j = 1) \tag{9}$$

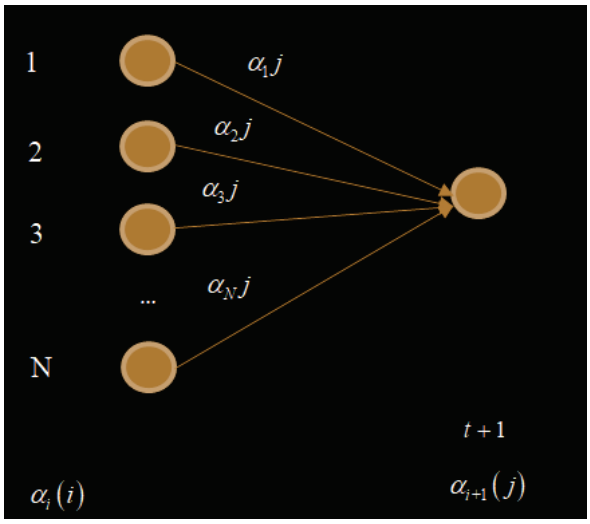

Figure 3 Diagram of forward probability.

Second, the recursive is taken:

$$\alpha_{t+1}(j) = \left[\sum_{i=1}^{N} \alpha_t(i) a_{ij}\right] b_j\left(o_{t+1}\right) 1 \leq i \leq N 1 \leq t \leq T-1 \tag{10}$$

Finally, the output probability is obtained:

$$p(O|\lambda) = \sum_{i=1}^{N} p\left(o_1, o_2, \cdots, o_t, \mathrm{s}_T = s_i|\lambda\right) = \sum_{i=1}^{N} \alpha_T(i) \tag{11}$$

Definition of backward probability:

$$\beta_t(i) = p\left(o_{t+1}, o_{t+2}, \cdots, o_T|s_t = s_i, \lambda\right) \tag{12}$$

Its meaning is the probability that the output vector is $o_{t+1}, o_{t+2}, \cdots, o_T$ from $t+1$ to T under the premise that the state at time t is i.

First, it is initialized:

$$\beta_T(N) = 1 \beta_T(j) = 0 (j = N) \tag{13}$$

Second, the recursive is taken:

$$\beta_t(j) = \sum_{j=1}^{N} a_{ij} b_j\left(o_{t+1}\right) \beta_{t+1}(j)\ 1 \leq i \leq N 1 \leq t \leq T-1 \tag{14}$$

Finally, the output probability is obtained:

$$p(O|\lambda) = \sum_{i=1}^{N} p\left(o_1, o_2, \cdots, o_T|\mathrm{s}_1 = s_i, \lambda\right) = \sum_{i=1}^{N} \beta_1(i) \tag{15}$$

The complexity of the calculation can be greatly reduced by the forward and backward algorithm.

When the observation vector O and the hidden Markov model parameter λ are given, the decoding problem requires determining the most likely state sequence of the output observation vector O from the model parameters λ. Specifically, if an HMM has M states and the time length is N, the number of all possible state sequences is M^N. Then, the decoding problem is to determine the most likely sequence of states, M^N, from which the observation vector 0 is generated.

The formula is described as:

$$s^* = \arg\max_{s} p(O, s|\lambda) \tag{16}$$

Then, this problem can be solved by the Viterbi algorithm. First, we need to define the function:

$$\delta_t(i) = \max_{\mathrm{s}_1, \mathrm{s}_2, \cdots \mathrm{s}_{t-1}} p\left(\mathrm{s}_1, \mathrm{s}_2, \cdots \mathrm{s}_{t-1}, s_t = \mathrm{i}, o_1, o_2, \cdots, o_t|\lambda\right) \tag{17}$$

The above equation shows the maximum probability path with state i at time t.

First, it is initialized:

$$\partial_1(i) = \pi_i b_i(o_1) \tag{18}$$

Second, the recursive is taken:

$$\delta_t(i) = \max_{1 \leq i \leq N} \left[\delta_{t-1}(i) a_{ij}\right] b_j\left(o_t\right) 2 \leq t \leq T 1 \leq j \leq N \tag{19}$$

Finally, the optimal sequence output probability is determined:

$$p^{\max} = \max_{1 \leq i \leq N} \left[\delta_T(i)\right] \tag{20}$$

The biggest difference between the Viterbi algorithm and the forward algorithm is the difference between the summation and the maximum value. The forward algorithm is used for summation and the Viterbi algorithm is used to find the maximum value. The output probability of the general optimal path will account for more than 99.5% of the output probability of all paths. The training problem is how to obtain the specific value of the HMM model parameters through a better algorithm when the observation sequence O is obtained. Moreover, that the likelihood probability $p(O|M)$ of the observation vector 0 is largest under this model parameter is satisfied.

Derivation of the EM algorithm:

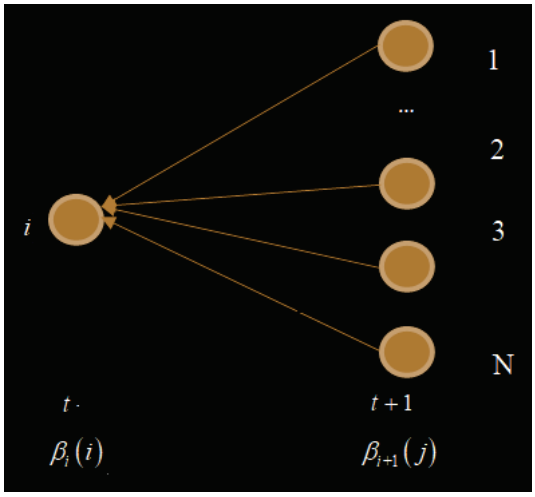

Figure 4 Diagram of backward probability.

(1) Define the function:

$$\zeta_t(i,j) = p\left(s_t = s_t, s_{t+1} = s_j | O, M\right) = \frac{a_t(i)a_{ij}b_j\left(o_{t+1}\right)\beta_{t+1}(j)}{\sum_{i=1}^{N}\sum_{j=1}^{N} a_t(i)a_{ij}b_j\left(o_{t+1}\right)\beta_{t+1}(j)} \quad (21)$$

This indicates the probability that the time t is the state i and the time $t+1$ is the state j under the premise that the model input parameters and observation vectors are given.

(2) Define the function:

$$\gamma_t(i) = \sum_{j=1}^{N}\zeta_t(i,j) \quad (22)$$

This indicates the probability that state t is the state i when the model parameters and the observation vector are given.

(3) Define the function:

$$\sum_{t=1}^{T-1}\gamma_t(i) \quad (23)$$

This indicates the probability that the vector is in state i in the case when the model parameters and the observation vector are given.

(4) Define the function:

$$\sum_{t=1}^{T-1}\zeta_t(i,j) \quad (24)$$

This indicates the probability of state i jumping to state j in the case where model parameters and observation vectors are given.

(5) Under the above definition, the model parameters are updated as follows:

$$***\pi_i = \gamma_t(i) \quad (25)$$

The equation above indicates the probability of being in state i at time $t = 1$.

$$\hat{a}_{ij} = \frac{\text{Probability of state i transitioning to state j}}{\text{Probability of state i}} = \frac{\sum_{t=1}^{T-1}\zeta_t(i,j)}{\sum_{t=1}^{T-1}\gamma_t(i)} \quad (26)$$

The formula above gives the weight update between nodes.

$$\hat{b}_i\left(o_k\right) = \frac{\text{Probability of sample } o_{\text{k}} \text{ in j state}}{\text{Probability of state j}} = \frac{\sum_{i=1}^{T}\gamma_i(j)}{\sum_{i=1}^{T}\gamma_i(i)} \quad (27)$$

The equation above indicates the likelihood of an update of the j state.

Assuming $M = \{A, B, \pi\}$ is the original model parameter and $** M = \{**\text{A}, **\text{B}, **\pi\}$ is the model parameter after revaluation, then the following conclusions can be proved:

1. When the model is already optimal, there must be:

$$M = ***M \quad (28)$$

2. The parameters of the new model make:

$$P(M/***M) > P(O/M) \quad (29)$$

It can be seen that as the iteration continues, $***M$ will converge to the optimal parameters.

3. MODEL BUILDING

As shown in Figure 5, the data input to the speech feature extraction module is waveform data obtained by pre-processing the simulated speech signal, and then sending it to

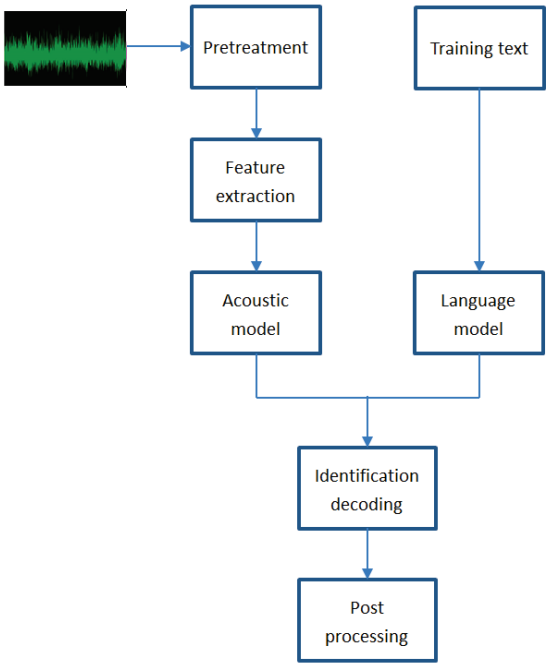

Figure 5 System framework for speech recognition.

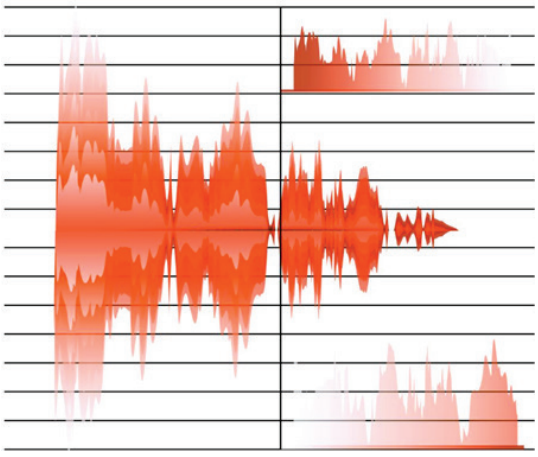

Figure 6 Spectrogram corresponding to the time domain signal waveform.

the feature extraction module. All speakers are different in terms of age, gender, pronunciation habits and feelings, some of which will change with time. Hence, people will produce different voice signals when expressing the same content. Moreover, to a large extent, acoustic features represent speech signals. Good acoustic characteristics should try to meet three conditions. Firstly, the acoustic characteristics should be well differentiated, the differences between the similarities should be as small as possible, and the differences between different classes should be as large as possible, which makes it easier to identify different information, and also facilitates more accurate modeling of different acoustic modeling units. Secondly, the extraction of speech features can be regarded as the compression coding process of speech information, and it is necessary to retain the information related to the speech content and eliminate other factors that are not closely related to the content. Moreover, it is necessary to reduce the dimension of the parameter when the information is retained enough, that is, the feature dimension should be moderate, so as to accurately and efficiently enter the training of the acoustic model. Finally, reliability and independence need to be considered, and there must be the ability to prevent interference from environmental noise.

(1) Spectrogram

A speech signal is usually analyzed in terms of time or frequency. However, both methods have limitations: Time domain analysis does not have an intuitive understanding of the frequency characteristics of speech signals, and frequency domain analysis cannot determine the relationship of speech over time.

The spectrogram is a map that expresses three-dimensional information on a two-dimensional plane. Its abscissa indicates time, the vertical scale indicates frequency, and the gray value of each pixel reflects the energy of the corresponding time and corresponding frequency. The spectrogram allows the properties of the phoneme to be well observed. The formant (the thicker bar in the spectrogram) carries the distinguish

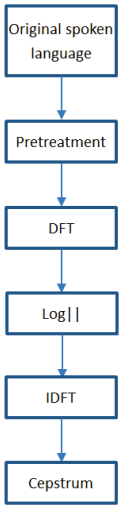

Figure 7 Flow chart for calculating the cepstrum coefficient

ability of the speech, and the speech can be better recognized by observing the variation of the formant.

(2) Feature Extraction

This is commonly used for speech recognition features and includes: Linear Prediction Coefficients (LPC), Mel-Frequency Cepstral Coefficients (MFCC), and Perceptual Linear Prediction (PLP). With linear predictive analysis, the speech signal at a certain moment can be linearly represented by the combination of signals at several previous moments. The basic problem is to directly derive a set of linear prediction coefficients based on the speech signal. When the mean square error (MSE) between the linear prediction estimate and the sampled value of the speech signal reaches a minimum value, the linear prediction coefficient can be extracted. The most important speech feature parameters are those extracted by means of cepstrum analysis, and the cepstral coefficients are implemented based on the homomorphic processing method that transforms a nonlinear problem into a linear problem. First, the original speech signal (actually a convolutional signal) is subjected to Discrete Fourier Transform (DFT) to obtain the spectrum (at this time, it becomes a multiplicative signal, and the convolution of the time domain is equivalent to the product of the frequency domain.). The discrete spectrum then takes the logarithm to turn the multiplicative signal into an additive signal. Finally, it is restored to a convolution signal by Inverse Discrete Fourier Transform (IDFT) to obtain cepstral coefficients. This method of finding cepstral coefficients can obtain relatively stable speech feature parameters.

4. ANALYSIS AND DISCUSSION

4.1 Speech Recognition Experiment

The Language Model (LM) is a customary way of describing human language. It mainly reflects the intrinsic relationship between words and words in the organizational structure, and the language model determines which word sequence is more likely, and several words are known to predict the next word. The language model that can accurately describe the law of language change directly affects the efficiency and performance of decoding, which is directly related to the overall performance of speech recognition. According to different production methods, language models can be grouped under two categories, namely, statistical-based language models and rule-based language models. The statistical-based language model is obtained by training a large amount of data. It describes the language model from the mathematical point of view and can calculate the probability of occurrence of each sentence in the original language. Rule-based language models require linguistic scholars and experts as they are based on linguistically relevant knowledge and are artificially compiled in conjunction with real-life language situations. However, this language model has limitations when dealing with large-scale real text, so it is generally not used.

In order to verify the validity of this model, the recognition rates of this model and other models in speaker-independent isolated word recognition are compared through experiments. In this paper, we use the Spoken ArabicDigit dataset in the UCI machine learning library, which includes 8800 Arabic digital voice data (88 people pronounce 10 Arabic numbers, repeating each one ten times) 6600 pronunciations of the first 66 people as training set, and 2200 pronunciations of 22 people as test sets. The software used for the experiment is MATLAB R2013a.

Firstly, the speech feature parameters are averaged into segments, and the speech feature parameters can be expressed as S (K, J), where K is the order of the feature parameters, J is the number of frames of the feature parameters after segmentation, and T is the number of original speech frames. The formula for calculating the average characteristic parameters into N segments is:

$$M(i) = S(\mathrm{K}, \mathrm{J}), \mathrm{J} = \left[\frac{T}{N}(\mathrm{i}-1)+1\right], \ldots, \left[\frac{T}{N}i\right] \quad (30)$$

The total number of speech feature parameters at each stage is shown in Table 1. The data given in Table 1 indicates that the piecewise mean dimension reduction regularization

Table 1 Dimension Reduction and Regularization of Speech Feature Parameters by Segmental Means.

factors	I	II	III	IV	V	VI
Matrix size	$T \times K$	$\left(\frac{T}{N}\right) \times K$	$\left(\frac{T}{NM}\right) \times K$	$\left(\frac{T}{NM} \times \frac{1}{\frac{T}{NM}}\right) \times K$	$M \times K$	$M \times N \times K$
Number of parameters	$T \times K$	$T \times K$	$T \times K$	$K \times M \times N$	$K \times M \times N$	$K \times M \times N$

Table 2 Evaluation Index Results-Sample Number.

index	Consistent	One level difference	Two level difference	Three level difference
Pitch	207	32	1	0
Speech rate	197	43	0	0
Rhythm	204	33	3	0
Tone	192	44	4	0

Table 3 Evaluation results of experimental indicators.

index	Consistency rate	Adjacent agreement rate	Pearson
Pitch	86.25%	99.58%	0.8
Speech rate	82.08%	100%	0.493
Rhythm	85.00%	98.75%	0.543
Tone	80.00%	98.33%	0.627

algorithm can reduce the dimension of the characteristic parameter matrix with the size of T $*$ K to the parameter matrix with the size of K $*$ M $*$ N. This algorithm successfully removes the influence of the number of speech frames T on the size of the data after dimensionality reduction regularization. The size of the parameter matrix after dimensionality reduction regularization is related only to the order K of the characteristic parameters, the size N of the segment and the size M of the sub-segment, which means that speech of different lengths can be regularized into the same size matrix, thereby greatly facilitating the application and improvement of the speech recognition algorithm.

Choosing the right acoustic model modeling unit is essential and the first problem encountered in acoustic modeling. A suitable granularity modeling unit can significantly improve the performance of the speech recognition system. A good acoustic modeling unit should have three attributes: consistency, trainability, and sharing. Consistency means that the same modeling unit in different speech instances requires the basic consistency of acoustic pronunciation. Trainability means that each modeling unit can correspond to enough modeling instances. Sharing means that the practice of different modeling units can share common training examples.

4.2 Experimental Results and Analysis

According to the method described in this paper, the scores for four indicators-pronunciation, speed, rhythm and intonation– in 240 sentences (10 sentences for each of the 24 students) can be obtained. The results are compared with those obtained through manual evaluation. The experimental results are shown in Tables 2 and 3.

In terms of intonation, 207 samples are evaluated with the same level of machine evaluation and manual evaluation, and the number of samples with a level of difference is 32. The difference between two grades is only 1, and there is no difference in the number of samples of three levels. It shows that the rate of agreement between machine and artificial pitch is 86.25% , the adjacent concordance rate is as high as 99.58%, and the Pearson correlation coefficient is 0.8, indicating that the intonation evaluation method is feasible.

In terms of speech speed, 197 samples are evaluated with the same level of machine evaluation and manual evaluation. The number of samples with the difference level is 43, and there is no difference between two or three samples. This shows that the rate of agreement between machine and artificial speech speed is 82.08% , the adjacent concordance rate is 100% , and the correlation coefficient is 0.493, indicating that the speed evaluation method is credible.

In terms of rhythm evaluation, there are 204 samples with the same level of machine evaluation and manual evaluation. The number of samples with the difference level is 33, the difference between two grades is only 3, there is no difference of three level samples, indicating that the rate of consistency between machine and artificial rhythm is 85% , the adjacent concordance rate is as high as 98.75% , and the Pearson correlation coefficient is 0.543, indicating that the rhythm evaluation method in this paper is credible. Some of experimental results are shown in Table 4.

Acoustic modeling units commonly used in speech recognition have syllables, phones, and tri-phones. As a modeling unit, the consistency of syllables is weak, but its training is strong, and can be applied to digital string recognition scenarios. As a modeling unit, the consistency of phonemes is general, but its trainability is relatively strong and suitable for isolated word recognition application scenarios. The ternary phoneme is very consistent as a modeling unit, but it is weakly trainable and suitable only for large vocabulary and large-scale speech recognition applications. In speech recognition, the phenomenon of co-pronunciation is considered, that is, each

Table 4 Experimental Data.

index		Pitch	speed	rhythm	tone	overall
01	Machine rating	A	A	B	B	A
	Manual rating	A	A	B	B	A
02	Machine rating	A	A	B	B	A
	Manual rating	A	A	B	B	A
03	Machine rating	A	A	B	B	A
	Manual rating	A	B	B	C	A
11	Machine rating	A	A	B	B	A
	Manual rating	A	A	B	B	A
12	Machine rating	A	A	B	A	A
	Manual rating	A	A	B	A	A

pronunciation may be distorted by the influence of adjacent sounds, and the context-dependent acoustic unit is usually selected as the modeling unit of the acoustic model. If only the influence of the previous note on the current note is considered, it is called a bi-phone. If both the previous note and the influence of the next note on the current note are considered, it is called a tri-phone. In this paper, context-sensitive triphones are used as the modeling unit in English speech recognition.

Unlike cepstral properties and linear prediction, the Mel cepstral coefficients and perceptual linear predictions are based, to some extent, on the mechanism of human auditory perception. The MFCC process involves first converting the signal from the time domain to the frequency domain by FFT, and then convolving the logarithmic energy spectrum with a set of uniform triangular filters in the Mel frequency domain. Finally, the output of the filter bank is converted by the discrete cosine transform method, and then several coefficients after DCT are taken as MFCC. At this time, a series of cepstrum vectors can be used to describe the speech, and each vector is the MFCC feature vector of each frame. Numerous studies have shown that MFCC parameters are superior to other parameters for the improvement of speech recognition systems in terms of performance.

5. CONCLUSION

Speech recognition is a very representative cutting-edge technology in the field of artificial intelligence, which is directly related to the future life experiences of human beings. The combination of deep learning and speech recognition will definitely advance the field of artificial intelligence. This study constructed an English speech recognition system based on and from the perspective of machine learning. Furthermore, it tested and analyzed the performance of the system, which will contribute to the development of artificial intelligence. At the same time, the subject was based on deep learning, and firstly studied the acoustic modeling based on DNN, and introduced the network structure and algorithm of DNN. Then, the DNN-HMM-based acoustic model was trained with 300 hours and 500 hours of English speech data, respectively. In addition, this paper optimized the English language model by means of the sub-word modeling method, which alleviates the problem of sparseness and robustness of the traditional whole-word language model brought by the super-large vocabulary of adhesive words.

REFERENCES

1. Bleisch S., Duckham M., Galton A., Laube P. (2014). Mining candidate causal relationships in movement patterns. International Journal of Geographical Information Science, 28 (2), 363–382.
2. Cao Yu-lin, Wang Xiao-ming, He Zao-bo, (2016). Optimal Security Strategy for Malware Propagation in Mobile Wireless Sensor Networks. Acta Electronica Sinica, 44(8), 1851–1857
3. Dorothy N D, AdrienneR, Neuman A C. (2017). Speech Recognition in Nonnative versus Native English-Speaking College Students in a Virtual Classroom, Journal of the American Academy of Audiology, 28(5):404–414.
4. Jesse A, Poellmann K, Kong Y Y. (2017). English Listeners Use Suprasegmental Cues to Lexical Stress Early During Spoken-Word Recognition, Journal of Speech Language and Hearing Research, 60(1):1–9.
5. Kang L, Liheng X, Jun Z. (2015). Co-Extracting Opinion Targets and Opinion Words from Online Reviews Based on the Word Alignment Model, IEEE Transactions on Knowledge and Data Engineering, 27(3):636–650.
6. Liang J J, Qu B Y, Mao X B, Et Al., (2014). Differential Evolution Based on Fitness Euclidean-distance Ratio for Multimodal Optimization. Neurocomputing, 137(8), Pp. 252–260.
7. Mushtaq M., Akram U., Khan I., Khan S. (2017). Cloud computing environment and security challenges: A review. International Journal of Advanced Computer Science and Applications, 8 (10), 183–195.
8. Mishra D., Samantaray S., Joos G. (2016). A combined wavelet and data-mining based intelligent protection scheme for microgrid. IEEE Transactions on Smart Grid, 7 (5), 2295–2304.
9. Mernik M, Liu S H, Karaboga M D, et al., (2015). On Clarifying Misconceptions When Comparing Variants of The Artificial Bee Colony Algorithm by Offering A New Implementation. Information Sciences, 291(10), 115–127.
10. Mahboubi H, (2014). Distributed Deployment Algorithms for Efficient Coverage in A Network of Mobile Sensors with Nonidentical Sensing Capabilities. IEEE Transactions on Vehicular Technology, 63(8), 3998–4016.
11. Meltzner G S, Heaton J T, Deng Y, et al. Silent Speech Recognition as an Alternative Communication Device for Persons with Laryngectomy, IEEE / ACM Transactions on Audio, Speech, and Language Processing, 2017, 25(12):2386–2398.
12. Stoean C, Preuss M, Stoean R, et al., (2010). Multimodal Optimization by Means of A Topological Species Conservation Algorithm. IEEE Transactionson Evolutionary Computation, 14(6), 842–864.

13. Sun Chao, Yang Chunxi, Fan Sha, et al., (2015). Energy Efficient Distributed Clustering Consensus Filtering Algorithm for Wireless Sensor Networks. Information and Control, 44(3), 379–384.
14. Upadhyay N, Rosales H. (2018). Robust Recognition of English Speech in Noisy Environments Using Frequency Warped Signal Processing, National Academy Science Letters, 20,110–112.
15. Wang X., Song Y. (2018). Uncertainty measure in evidence theory with its applications. Applied Intelligence, 48 (7), 1672–1688.
16. Wang Jie, Lu Jianzhu, Zeng Xiaofei, (2016). Data Aggregation Scheme for Wireless Sensor Network to Timely Determine Compromised Nodes. Journal of Computer Applications, 36(9), 2432–2437.
17. Zhang R., Gao M., He X., Zhou A. (2016). Learning user credibility for product ranking. Knowledge and Information Systems, 46, 679–705.