Hindawi
Complexity
Volume 2023, Article ID 9875079, 1 page
https://doi.org/10.1155/2023/9875079

Retraction

Retracted: Emotional Interactive Simulation System of English Speech Recognition in Virtual Context

Complexity

Received 22 August 2023; Accepted 22 August 2023; Published 23 August 2023

This article has been retracted by Hindawi following an investigation undertaken by the publisher [1]. This investigation has uncovered evidence of one or more of the following indicators of systematic manipulation of the publication process:

(1) Discrepancies in scope
(2) Discrepancies in the description of the research reported
(3) Discrepancies between the availability of data and the research described
(4) Inappropriate citations
(5) Incoherent, meaningless and/or irrelevant content included in the article
(6) Peer-review manipulation

The presence of these indicators undermines our confidence in the integrity of the article's content and we cannot, therefore, vouch for its reliability. Please note that this notice is intended solely to alert readers that the content of this article is unreliable. We have not investigated whether authors were aware of or involved in the systematic manipulation of the publication process.

Wiley and Hindawi regrets that the usual quality checks did not identify these issues before publication and have since put additional measures in place to safeguard research integrity.

We wish to credit our own Research Integrity and Research Publishing teams and anonymous and named external researchers and research integrity experts for contributing to this investigation.

The corresponding author, as the representative of all authors, has been given the opportunity to register their agreement or disagreement to this retraction. We have kept a record of any response received.

References

[1] D. Li, "Emotional Interactive Simulation System of English Speech Recognition in Virtual Context," *Complexity*, vol. 2020, Article ID 9409630, 11 pages, 2020.

Hindawi
Complexity
Volume 2020, Article ID 9409630, 11 pages
https://doi.org/10.1155/2020/9409630

Research Article

Emotional Interactive Simulation System of English Speech Recognition in Virtual Context

Dan Li

School of Foreign Languages, Luoyang Institute of Science and Technology, Luoyang, Henan 471023, China

Correspondence should be addressed to Dan Li; 200901200802@lit.edu.cn

Received 25 May 2020; Accepted 27 July 2020; Published 11 August 2020

Guest Editor: Zhihan Lv

With the development of virtual scenes, the degree of simulation and functions of virtual reality have been very complete, providing a new platform and perspective for teaching design. Firstly, the hidden Markov chain model is used to perform emotion recognition on English speech signals. English speech emotion recognition and speech semantic recognition are essentially the same. Hidden Markov style has been widely used in English speech semantic recognition. The experiments of feature extraction and pattern recognition of speech samples prove that Hidden Markovian has higher recognition rate and better recognition effect in speech emotion recognition. Secondly, combining the human pronunciation model and the hearing model, by analyzing the impact of the glottis feature on the human ear hearing-model feature, the research application of the English speech recognition emotion interactive simulation system uses the glottis feature to compensate the human ear, hearing feature is proposed by compensated English speech recognition, and emotion interaction simulation system is used in the English speech emotion experiment, which has obtained a high recognition rate and showed excellent performance.

1. Introduction

Situation, virtual situation and instructional design are the starting points of this thesis. The teaching research has been concentrated on the discussion of teaching content and teaching methods, and there is almost no teaching research on the perspective and level of time and space and artistic conception. The research on the situation is also focused on the research of the "situation teaching" or "situation learning" theory and its teaching methods, or the pure situation or concept research is very fragmented, unsystematic, and not combined with educational discussions. In order to deepen the situation-based teaching research, especially the English speech recognition emotion interactive simulation teaching design in virtual reality, it is necessary to study and explore the essence of the situation and virtual situation and significance of teaching, which seems very meaningful.

Usually speech recognition technology includes three parts: preprocessing, feature extraction, and pattern matching [1, 2]. In the preprocessing process, the prefiltering, pre-emphasis, windowing, and framing of the original voice signal are completed. Feature extraction is one of the key issues of speech recognition technology. In this link, the preprocessed speech will be mapped to the feature space, and the mapping function is the feature extraction algorithm [3, 4]. Another key issue of speech recognition technology is pattern matching, which divides the extracted features into two parts, called training features and test features, respectively. Substitute the training feature into the pattern matching algorithm, train one or several models, and then input the test feature into the trained model to find the model with the highest degree of matching; then, the model is considered to be the representative of the test feature category. The best combination of feature extraction and pattern matching will produce the best recognition effect. Speech synthesis systems are also called text-to-speech conversion systems. As the name implies, they are to convert text into speech [5, 6], which corresponds to speech recognition technology. The purpose of speech synthesis technology is simply to let the computer speak human-like language. In a speech synthesis system, the grammatical

ordering of basic phonemes, that is, the definition of annotated sequences, is the key to synthesizing grammatical rules with high intelligibility [7, 8]. At present, most definitions for labeling sequences are based on the relationship between the current phoneme and the current sentence. If you define the relationship between the front and back sentences and the front and back phonemes and the current phoneme in detail and use these relationships to generate synthesized speech in the synthesis stage, then it is inevitable. It will make the grammatical ordering of phonemes more in line with language habits, resulting in synthetic sentences with higher intelligibility and naturalness [9]. The literature [10, 11] makes a comprehensive and detailed analysis of this theory from the background, knowledge, contextuality of cognition and learning, assumptions, and basic characteristics of contextual learning and contextual cognition. The literature [12, 13] reviews the research of situational cognition and learning theory in the West, summarizes the research history, the meaning, and main content of situational cognition and learning, and points out the future development of the theory direction. The literature [14, 15] conducted many experiments and innovations on Chinese speech emotion recognition. Exploratory experiments on speech emotion recognition include the use of various improved neural network methods, support vector machines, and fuzzy entropy recognition methods. The test results can achieve a recognition rate up to about 75%. The authors in [16, 17] have also conducted fruitful research on feature vector decomposition and fusion of speech emotion signals and feature vector dimensionality reduction, proposed improved classification quadratic functions, and improved Mahalanobis distance and other classification algorithms, and recognized in experimental data the good test performance on the rate. Currently, commonly used speech synthesis methods are parameter synthesis method and waveform stitching synthesis method [18, 19]. The parameter synthesis method is to synthesize speech by adjusting acoustic parameters, which can be simply understood as the reverse process of speech recognition. The parameter synthesis method is relatively mature in theory, but because of the complexity of the algorithm and because the loss of information cannot be effectively compensated, the synthesis effect is not ideal and has a strong machine smell. In the waveform splicing and synthesis system, the basic phonemes are selected from a prerecorded speech database. As long as the database is large enough, it is theoretically possible to splice out any sentences needed [20, 21]. Since the synthesized speech is basic the phonemes are all from the natural original pronunciation of the recorder, so the clarity and naturalness of the synthesized sentences are relatively high. However, it requires high storage capacity and operation speed. In recent years, the HMM-based speech synthesis method has attracted widespread attention. The HMM-based synthesis system can solve the defects of the waveform stitching synthesis method. It models the speech parameters, then uses the sound database data for automatic training, and finally forms a corresponding synthesis system [22]. Emotional speech recognition is an important branch of speech recognition, which has aroused widespread concern in recent years. Similar to speech recognition, emotional speech recognition can also be divided into three parts: preprocessing, feature extraction, and pattern matching [23, 24]. The difference is that the features here refer specifically to those features that can characterize the emotional state, such as the pitch Rhythm characteristics such as frequency, formant, and short-term average energy. Rhythm feature extraction algorithm is simple, but it is not effective enough for some states with less emotional opposition [25, 26]. Studies [27, 28] in recent years have shown that emotional speech features are no longer a simple combination of the above prosody features. Features based on pronunciation mechanisms are increasingly used in emotional speech recognition and have achieved good results.

The in-depth analysis and discussion of the situation and the virtual situation provide a more in-depth theoretical basis for the development and teaching application of constructivism, provide a more practical plan, and open a new perspective and thinking for the teaching design in virtual reality. First, it introduces in detail the establishment, acquisition, and storage methods of the English speech recognition emotion interaction emotion speech sample library, classifies it according to the requirements of subsequent experiments, and then explains how to affect the preprocessed speech emotion signal. Next, we introduce and compare the main English speech recognition emotion interaction speech emotion feature parameters currently used and how to use the experimental method to obtain the feature parameter fields required in this article. Second is an introduction to the process of speech emotion recognition, including the current major voice emotion recognition methods in the world, comparing their respective characteristics, and then choosing the hidden Markov model method as the voice emotion recognition method in this paper to complete the recognition of virtual human emotion recognition of English speech recognition emotion interactive speech emotion signals. Find the Hidden Markov Model that is most suitable for the purpose of this article, conduct a detailed comparison experiment between discrete Hidden Markov and continuous Hidden Markov Models, draw a comparison of the two advantages and disadvantages, and finally choose continuous Hidden Markov. The model serves as a model for further work.

2. Virtual Situation and Emotion Theory of English Speech Recognition

2.1. Design of English Phonetics Teaching Based on Virtual Situation. Virtual context provides users with a set of experiences, which is an external experience compared with the internal experiences of people's inner activities (such as joy, anger, sorrow, and joy). Because any external experience may cause people's inner activities, virtual situations can also cause people's inner activities. Therefore, the virtual situation can enable the learner to produce the same or similar inner activities as the real situation. Virtual situation is a phenomenon that can be directly observed and experienced, but it does not have the substantial accessibility as the real

world. Functionally speaking, the experience in the human mind is caused by the virtual situation and the real situation is not direct. When the users in the virtual situation conduct research and reflection, the virtual situation is a possible real reality. Since virtual situations are possible real realities, you can use virtual situations to understand real realities. Here, the virtual situation becomes a virtual intermediary.

The practice of English speech recognition emotional interaction teaching in virtual situations is in line with the constructivist theory and includes four basic attributes of "situation," "collaboration," "conversation," and "meaning construction." Constructivism believes that personal experience can accelerate the process of knowledge construction and realize the consolidation and externalization of knowledge, and the most convenient virtual situation is the most effective. The virtual situation allows the learner to interact with various existing information, experience different time and space in the learning process, and also be in contact with various parts of the virtual realm. The virtual situation is almost the same as the real situation in terms of simulation degree and function, which also meets the requirements of constructivism that the situation is "highly realistic." And when participating in the virtual situation, through the avatar mode, it can actively interact with other things and characters, participate in various English speech recognition emotional interactive teaching activities, and conduct collaborative learning to obtain knowledge and achieve the purpose of knowledge construction, as shown in Figure 1.

In virtual situation-based teaching, the learner plays the role of a cognitive subject. Therefore, in virtual situations, English speech recognition is often used for emotional interaction. English speech recognition emotional interaction refers to the support and promotion of learners' effective learning. Its core is to give full play to the initiative and enthusiasm of students in learning and fully reflect the learner's cognitive subjective role. Under constructivist thinking, many methods of situation-based teaching have been proposed. Learning in the context of virtual teaching can also draw on these methods. These methods highlight two aspects of collaboration and self-exploration. Constructivists are those who collaborate between teachers and students and between students and students, which is very important for the collection and analysis of data, the formulation and verification of hypotheses, the self-feedback of the learning process, the evaluation of learning results, and the construction of meaning. In the teaching based on the virtual teaching situation, discussion and communication are the main forms of collaboration, but the form of discussion and communication is carried out under the virtual situation. In the process of collaborative learning, teachers first ask questions to arouse students' "thinking and discussion." During the discussion, teachers further lead the problems deeper to deepen students' understanding and at the same time guide students to correct errors and supplement one-sided cognition.

2.2. Introduction to the Emotional State of Virtual Situations. According to the basic sentiment theory, the human emotional state can be divided into several basic types. Let the emotional state space set $T = \{t_i \mid i = 1, 2, \ldots, N\}$, where N represents the number of emotional states. Use random variables to scroll through emotional states. Let Q_i be the probability of $X = t_i$ (take the ith emotional state) and satisfy

$$\sum_{i=1}^{N} t_i = 1, \quad 0 \le t_i \le 1. \tag{1}$$

In this way, the probability space of emotional states can be expressed as follows:

$$\begin{pmatrix} T \\ Q \end{pmatrix} = \begin{pmatrix} t_1, t_2, \ldots, t_N \\ q_1, q_2, \ldots, q_N \end{pmatrix}. \tag{2}$$

In English speech emotion speech recognition, some features need to be found, and these features can accurately distinguish different emotional states. This issue involves a variety of disciplines such as psychology, biology, and signal processing. Table 1 compares the results of different acoustic characteristic parameters under six emotional states ("happy," "angry," "fear," "sadness," "amaze," and "disgust").

With the deepening of the research on the classification of English speech emotion interaction, some researchers have proposed a continuous English speech emotion interaction model, which describes emotion in a continuous space, that is, the dimension theory of emotion. Dimension theory regards the conversion process between different emotional states as a continuous linear process in an N-dimensional space, and the distance between different emotional states in the dimensional space represents the similarity and difference between them. Among the dimensional theory of emotion, the most widely accepted dimensional model is the two-dimensional space of activation evaluation:

(1) The degree of activation or arousal reflects the active degree of body energy in a certain emotional state

(2) The degree of evaluation or happiness is based on the separation and activation of positive and negative emotions

Based on the theory of activation evaluation space two-dimensional English speech emotion interaction classification model, as shown in Figure 2, each emotion is distributed at the outer end, and the center of the emotion wheel model is the natural origin, which is a comprehensive. The state of emotion tends to all emotions in the direction, and no emotion is reflected. The closeness of emotions in the same direction determines the intensity of each emotion, and the absolute value of the length of emotion shows the intensity of emotion in this aspect. The emergence of the English speech emotion interaction model introduces the vectorization theory into the classification of emotion models. Any emotion can be represented by a unique vector. According to the vectorization theory, naturally, the amplitude value of the emotion vector is expressed. The intensity of emotion and the angle of the emotion vector show the trend of the type of emotional interaction of English speech in this direction.

Introduce the continuous emotion model from the two-dimensional space to the secondary derived emotional

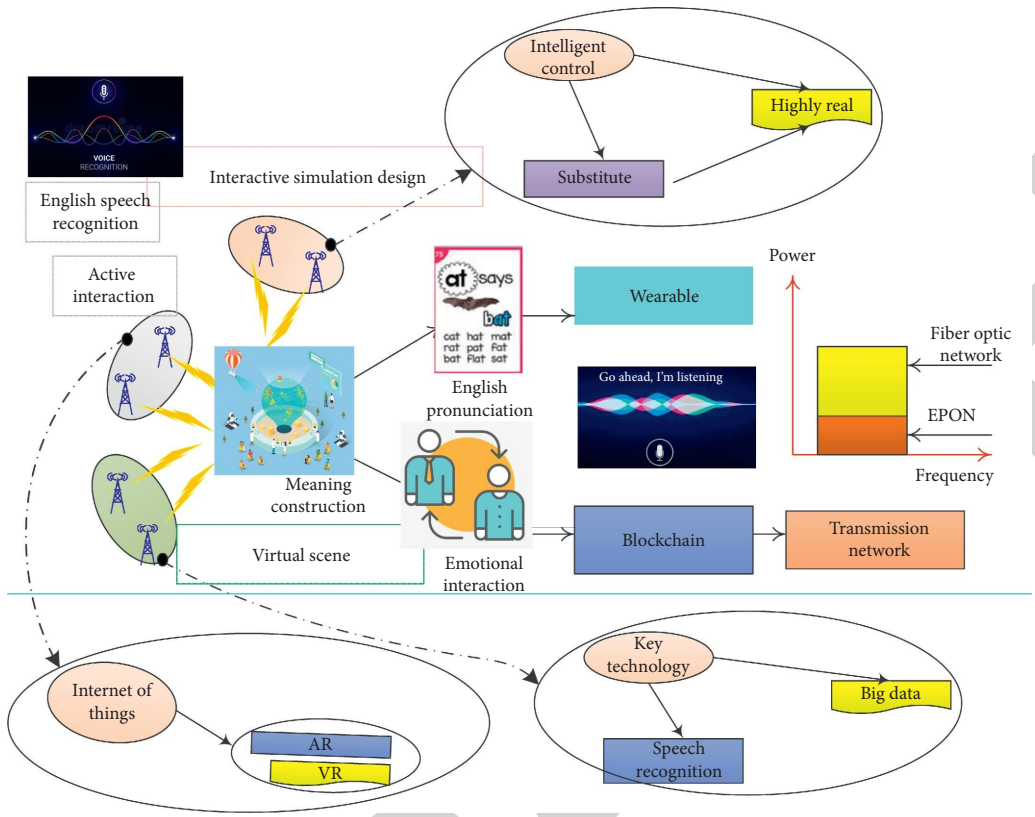

Figure 1: English speech recognition emotion interaction mode under virtual scenarios.

Table 1: Emotional and speech emotional features.

Parameter	Happy	Pissed off	Fear	Sad	Amazed	Disgust
Speed of speech	Fast	Slightly faster	Quickly	Slightly slower	Very fast	Very fast
Average pitch	Very high	Very high	Very high	Very low	Very high	Very low
Pitch range	Very wide	Very wide	Slightly narrower	Slightly narrower	Tip up	Slightly wider
Pitch change	Smooth	Mutation at stress	Bend down	Bend down	Normal	Normal
Strength	Bend up	High	Low	Low	Irregular sound	Normal
Sound quality	High	Breathing	Resonate	Resonate	Accurate	Irregular sound
Sharpness	Breathing	Chest sound	Vague	Vague	Amazed	Clear

factors, thus introducing the two-dimensional plane space into the three-dimensional space. More importantly, the emotion model in the three-dimensional space combines matrix theory and feature vector. The theory provides the basis of the emotion model for the subsequent vectorized decomposition and fusion of speech emotion feature parameters and the vectorized description of complex emotions in continuous space.

3. Research on Emotional Interaction of English Speech Recognition Based on Virtual Situation

3.1. English Speech Recognition Preprocessing. In order to convert the continuous analog voice signal into a digital signal for processing and recognition, before extracting the features of the English voice signal, it must go through a series of steps of sampling, digitization, and signal optimization to denoise the voice signal and select recognition frequency range, high-pass and low-pass filtering, endpoint detection, pre-emphasis, windowing, amplification and control gain, and antialiasing filtering and other voice signal specifications and preparations to facilitate the extraction of voice emotional feature parameters; these tasks are called preprocessing of emotional speech signals.

The English speech signal can be regarded as a non-stationary time-varying signal. Biological research on the principle of sound production by the vocal organs proves that the speed of the state of the vocal organs caused by vibration is much lower than the vibration frequency of the sound. Researchers of emotional signal processing in English speech recognition usually treat emotional signals in English speech recognition as short-term stable signals. In the sound frequency range of 5–50 ms, the spectral characteristics and some physical characteristics of English speech recognition emotion signals are constant. Based on this principle, we have introduced a "windowing" preprocessing method for

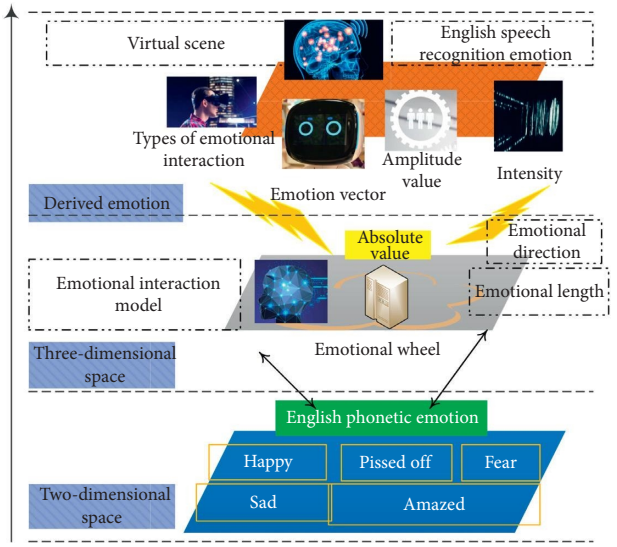

Figure 2: English speech emotion interaction classification model.

the English speech recognition emotion signal, combining the short-term processing method of the English speech recognition emotion signal with the processing method and theory of the stable process in a short time slice, and each short-term speech signal segment in a time slice is called an analysis frame. The length of a frame ranges from 10 ms to 30 ms. The "windowing" mentioned by the researchers is a short-term processing method for analyzing frames. The method is the English speech recognition emotion signal in a valid domain which is divided into many short time slices by artificially adding some window functions. The window function of the method of $w(n)$ to obtain the current frame is to change the sample amplitude outside the divided extraction processing area to zero. The most ideal state of the frequency response of the window function is no side-lobe spectrum leakage, the length is 0, the main-lobe spectrum leakage is close to no, and the length is infinitely narrow. This is only the ideal state that exists in the achieved simulation experiment. The two most commonly used window functions in the rectangular window and Hamming window English speech recognition emotional signal digital processing are

$$w(n) = \begin{cases} 1, & 0 \le n \le N-1, \\ 0, & n = \text{else}, \end{cases}$$
$$w(n) = \begin{cases} 0.58 - 0.42\left(\dfrac{2\pi n}{N-1}\right), & \\ 0, & \text{else}. \end{cases} \quad (3)$$

The characteristic parameters of the speech generated by the linear excitation source filter are mainly reflected in the frequency spectrum structure of speech, in which the formant and pitch frequency are the more commonly used and important characteristic parameters. Formants are one of the most important parameters in speech signal processing. The difference in the emotional state is reflected in the nervous system to form different degrees of nerve tension, and this difference in tension acts on the vocal muscles, making the same tone. The difference in the degree of muscle tightness is further reflected in the difference in channel frequency. This difference is reflected in the characteristic parameters of the speech signal processing process, which is the formant peak value. A large number of studies have proved that different emotional states will make the position of the formant of the speech signal different. In terms of statistical significance, in order to eliminate the individual differences in formants in the same emotional state on a statistical level, researchers widely use the statistical characteristics of the first three formants peak and bandwidth: average, extreme value, standard deviation, and median, as a research parameter in the process of speech signal processing.

In the extraction of feature parameters of speech signal processing, the linear prediction method based on the all-pole model regards the speech signal $x(n)$ as a full-pole filter response with $u(n)$ as the excitation, and from $u(n)$ to $x(n)$, it is expressed as follows:

$$T(z) = \frac{D}{1 - \sum_{i=1}^{q} b_i z^i}. \quad (4)$$

The linear prediction method is used to process the vocal tract model to produce an all-pole model with excellent performance in speech processing research. The practical significance of linear prediction is to decompress the speech signal in the time and frequency domains. The process of deconvolution is to treat the excitation component as the prediction residual $u(n) = x(n) - \sum_{i=1}^{q} b_i x(n-1)^i$ and obtain the component of the transfer function $T(z)$ of the full pole model. The parameter sequence a $[i]$ of this component is further obtained. According to the conclusion that the frequency response characteristic will reach a peak at the resonance peak frequency, the resonance peak is the spectral peak of the frequency response component at this time. By the formula,

$$T(z) = \frac{G}{1 - \sum_{i=1}^{q} b_i x(n-1)^i}. \quad (5)$$

Using the above method, the third formant-related characteristic data in Table 2 below is obtained. From the data in the table, we can think that the third formant has certain distinguishing effects on the four emotions.

For the purpose of feature parameter dimensionality reduction, this paper selects the first 10 numbers from the emotion package of the English speech recognition emotion library for a total of 70 speeches for 7 repetitive trainings, in order to obtain the optimal feature parameter vector combination. In the calculation process, there are two methods for selecting the best feature dimension, one is the cumulative contribution rate method, and the other is to obtain the recognition rate through repeated experiments, and the multiple experiment method of inferring the dimension based on the recognition rate. In the experiment process, only 70 voices were used for dimension reduction

Table 2: Statistical parameters related to different emotions of the third formant.

Third formant frequency	Sad	Happy	Fear	Pissed off
Mean	3048	3118	3005	3067
Maximum	4342	4067	4098	4089
Minimum value	2096	2106	1968	2234
Standard deviation	356	358	425	354
Median	1908	2926	2816	2945

and feature parameter selection, so the method using multiple experiments to find the dimension is more effective. In 7 repeated experiments, the statistical relationship between the average recognition rate and the dimensionality is shown in Figure 3.

As shown in Figure 3, it can be observed that when using feature vector groups with dimensions of 5 to 8, the recognition rate gradually increases; when the dimension exceeds 8, the recognition rate gradually decreases. Using 6 repeated experiments in this paper, the viewpoint of the best vector combination dimension is confirmed so that the conclusion based on the emotional speech library of this article is obtained. The 8-dimensional feature vector group can obtain the best recognition rate. The pitch average, pitch frequency first- and second-order difference, amplitude energy mean, amplitude energy first- and second-order difference, third formant standard deviation, and MFCC standard deviation are eight parameters as speech feature parameters.

3.2. English Speech Recognition Based on Virtual Scenarios. The hidden Markov model is used for emotion recognition. The model characteristics of the double embedded stochastic process characteristics are very consistent with the requirements of emotional state recognition through the emotional feature parameters in speech. The hidden Markov model method is widely used in the field of speech recognition and emotion recognition because of its many statistical advantages as a pattern recognition method; it can be more convenient to train approximate model parameters from a limited number of speech material data. Based on the flexibility of the trained model itself to change materials, optimize the architecture of the cognitive system as the number of training changes and improve the accuracy of the model.

Firstly, a continuous hidden Markov model is used to conduct speech recognition experiments. First, we need to initialize the emotional feature parameters of the speech. The pitch frequency, the first- and second-order difference numbers of the fundamental frequency, the amplitude energy value, the amplitude energy first- and second-order difference values, the first resonance peak, and the eight feature parameters, according to the degree of influence of each feature parameter plus weight value w_i (where $\sum_{i=1}^{8} w_i = 1$ represents the number of frames) to obtain the eight-dimensional feature vector, are shown below, where F_i represents the feature vector of each frame and i represents number of frames:

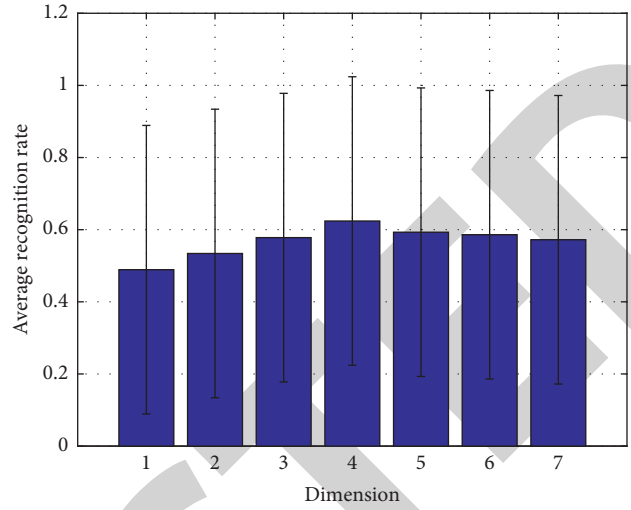

Figure 3: Relationship between average recognition rate and dimension.

$$F_i = \left(w_1 \times F_{0i}, w_2 \times \frac{dF_{0i}}{d_t}, \ldots, w_8 \times F_{1i} \right). \quad (6)$$

The iterative estimation algorithm is used to train the model of speech emotion parameters. The training results of this method are related to the initial value of the parameter (not the initial value of the system). Sometimes, due to the selection of training parameters, two types of situations will occur: the first type, the model cannot converge; in the second category, the converged solution is not the global optimal solution. In order to avoid and solve these two types of situations, we can use the piecewise K-means algorithm to make the training parameters converge better globally optimal solution:

(1) The input training voice feature parameters are segmented at equal intervals according to the number of states of the HMM model (that is, the number of classifications of the emotion model), and then the feature parameter set in each segment is used as a specific training set to train a certain Emotional state, calculating the initial parameters of the model.

(2) According to the incoming parameters, the preprocessed voice signal data is divided into the most likely state sequence.

(3) Use the K-means segmentation algorithm to iteratively re-evaluate the B in the initial parameters to obtain the intermediate result, that is, collect all the training data of the five emotional states obtained in step (2) to obtain B'; if it has been knowing that a model is divided into mixed Gaussian density functions with M mixed numbers, the K-means clustering algorithm is used to classify all training parameter sets with a state equal to l into M classes and the covariance matrix and mean value of each class. The vector is the center point of the class as the standard point for classification, and finally the mean

estimate M and variance estimate E of the M Gaussian components are regarded as the vector mean and co-square matrix of each type of feature vector set and each Gaussian component. The mixed weight value is as follows:

$$\vartheta_{jm} = \frac{\text{Number of voice frames in state } j \text{ in class } m}{\text{Number of voice frames in } j \text{ state}}. \quad (7)$$

(4) Substitute the λ obtained in the previous step into the parameter iterative re-estimation method as its initial value re-estimation model, and the re-estimation results in the new model parameter λ'.

(5) Find the difference between λ and λ', see if the difference is less than the preset threshold, and see if the template converges; if it does not converge, enter λ' as the new initial parameter and return to step (2). Go to step (6).

(6) Output the model parameter λ', that is, λ' is the final model parameter estimation result.

4. Experimental Verification

The selected English speech emotion recognition categories are sad, happy, fear, angry, and neutral. For these five emotional states, 162 groups of speech are selected for each emotion. A total of 810 groups of speech form a training speech database; divided into four times, the model is trained according to the model training volume of 210 groups, 410 groups, 610 groups, and 810 groups. The remaining 190 sets of corpus materials in the 1000 sets are used as the recognition set for the experiment, and the following two sets of tables are obtained.

As can be seen from Table 3, among the 200 groups of unfamiliar emotion speech recognition results, the recognition rate of angry emotion is the highest, reaching 81.2%, the recognition rate of sad emotion is the lowest, reaching 75.4%, and the other three emotions are happy, neutral, and fearful, whose recognition rates are 79.5%, 76.8%, and 77.8%, respectively, and the average recognition rate was 78.122%. Further analysis of the table also shows that the three emotions of sadness, fear, and neutrality are more easily confused in the recognition process. It is speculated that this situation is closely related to the emotional feature parameter domain of these three emotions.

As shown in Table 4, 190 sets of long-term speech and 190 sets of short-term speech were used to perform recognition experiments under 210, 410, 610, and 810 training amounts, respectively. The recognition rate results shown in the table below are obtained.

As shown in Table 4, the following conclusions are drawn. With 200 groups of untrained unfamiliar voices, the recognition rates are 52.53%, 61.96%, 72.49%, and 78.32% under the training volume of 210, 410, 610, and 810 groups, respectively. With the improvement of the training volume, the recognition performance of the resulting model is significantly improved. The recognition ability of the unfamiliar speech shows an upward trend as shown in Figure 4

Table 3: Emotion recognition results under 810 training amounts.

Emotion type	Sad	Happy	Pissed off	Fear	Neutral	Recognition rate
Sad	115	4	3	20	18	75.4
Happy	5	121	15	13	6	79.5
Pissed off	8	12	124	12	4	81.2
Fear	16	11	7	117	9	76.8
Neutral	14	3	4	20	119	77.8

Table 4: Recognition rate of 190 strange recognition corpora in different situations.

Training volume voice length	210	410	610	810
Long-term voice (3–6 seconds interval)	52.53%	61.96%	72.49%	78.32%
Short-term voice (2 seconds)	53.56%	65.79%	84.59%	89.87%

below. We can see that although increasing the training volume can recognize the performance, training in 610, the improvement rate of the recognition accuracy rate between the amount of training and the 810 groups of training is not as fast as the stage of 210 groups to 10 groups. Conclusion: increasing the amount of training can improve the accuracy of recognition. The accuracy improvement curve is infinitely close to the extreme value, and there is a bottleneck value in the system recognition rate.

The experimental recognition results of the 210 long-term speech recognition set under the 810 training set are shown in Table 5:

With the increase in training volume, the recognition performance of the resulting model has improved to a certain extent, but the overall recognition performance of the model for long-term speech in this paper is poor; while the model has a significantly higher recognition rate for short-term speech The recognition rate curve of long-term speech and short-term speech under different training amounts is shown in Figure 5:

As shown in Figure 6, the above five characteristics include not only classic characteristics reflecting the auditory characteristics of the human ear but also recently proposed characteristics reflecting the nonlinear characteristics of the glottis. Through these five characteristics of the recognition experiment of the English emotional speech database and the n emotional speech database, the comparison of the experimental results' data can directly prove that the English emotional speech database is a more effective emotional speech database, which can be used for future emotional speech research.

As shown in Figure 7, from the perspective of sentiment classification, the performance of LPCC, MFCC, and LPMCC features has changed. In a single speech database experiment, the "neutral" emotion recognition rate is the highest, "happy" is the second, and "angry" is the lowest.

Test in the same training corpus and rule set, observe the changes in the mark recall rate, and mark the accuracy rate. The results are shown in Figure 8.

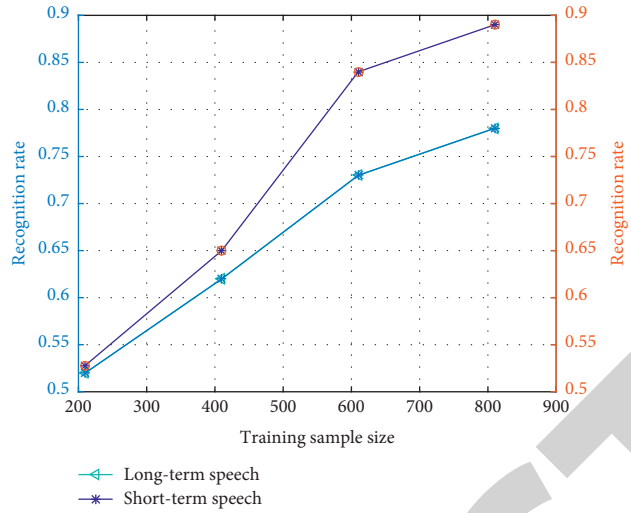

Figure 4: Recognition rate of 210 groups of strange voices under different training volumes.

Table 5: 210 long-term speech recognition results in the 810 training set.

Emotion type	Sad	Happy	Pissed off	Fear	Neutral	Unknown error	Recognition rate (%)
Sad	94	9	4	24	23	12	57.5
Happy	5	103	18	18	6	12	63.2
Pissed off	11	24	103	18	6	5	63
Fear	22	15	14	98	12	8	60.5
Neutral	30	11	9	19	88	4	55.8
Average recognition rate					59.995%		

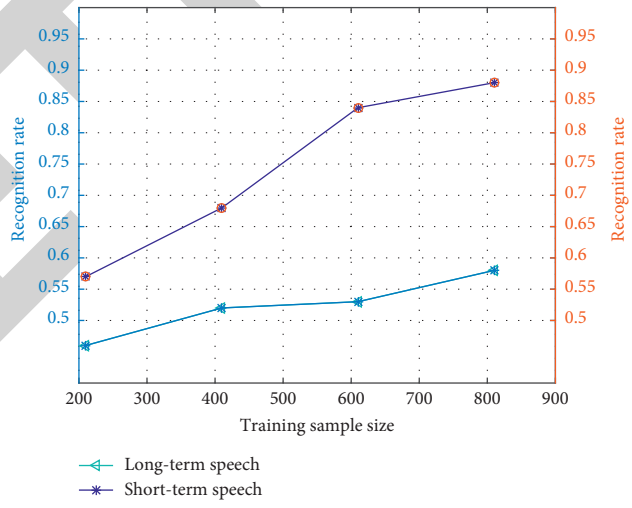

Figure 5: Long-term and short-term speech results under different training sets.

It can be seen from Figure 8 that when λ_1 takes 0, its LR value is relatively small, and the result is not ideal. Therefore, comprehensive consideration of the rule information and structure co-occurrence information is indeed very good for improving the accuracy of analysis great help. It can also be seen from the figure that when $\lambda_2 = 2.5$, the corresponding LP value is relatively high.

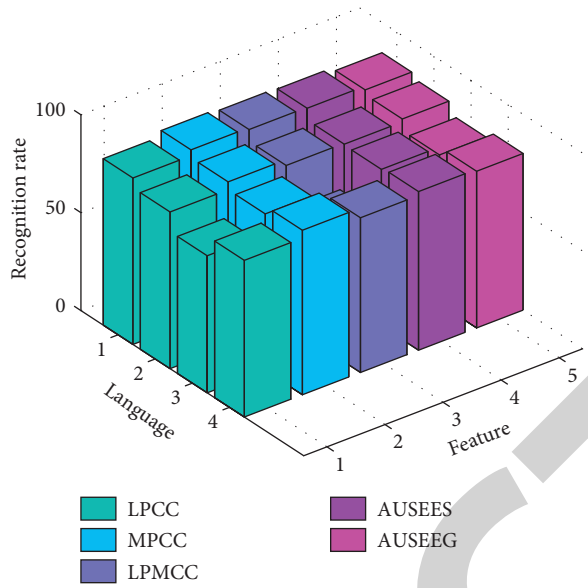

Figure 6: Comparison of recognition rates of different features in different languages.

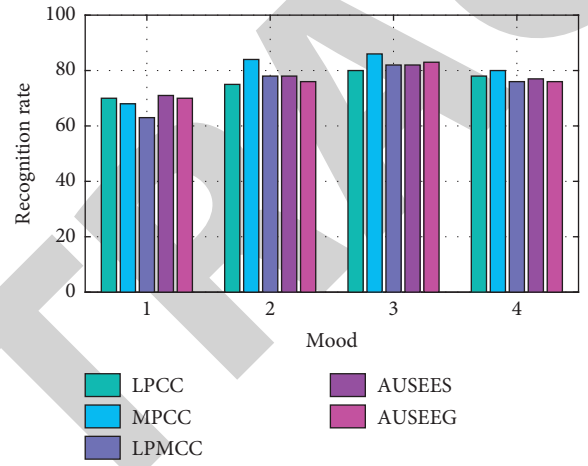

Figure 7: Comparison of recognition rates of different features under different emotions.

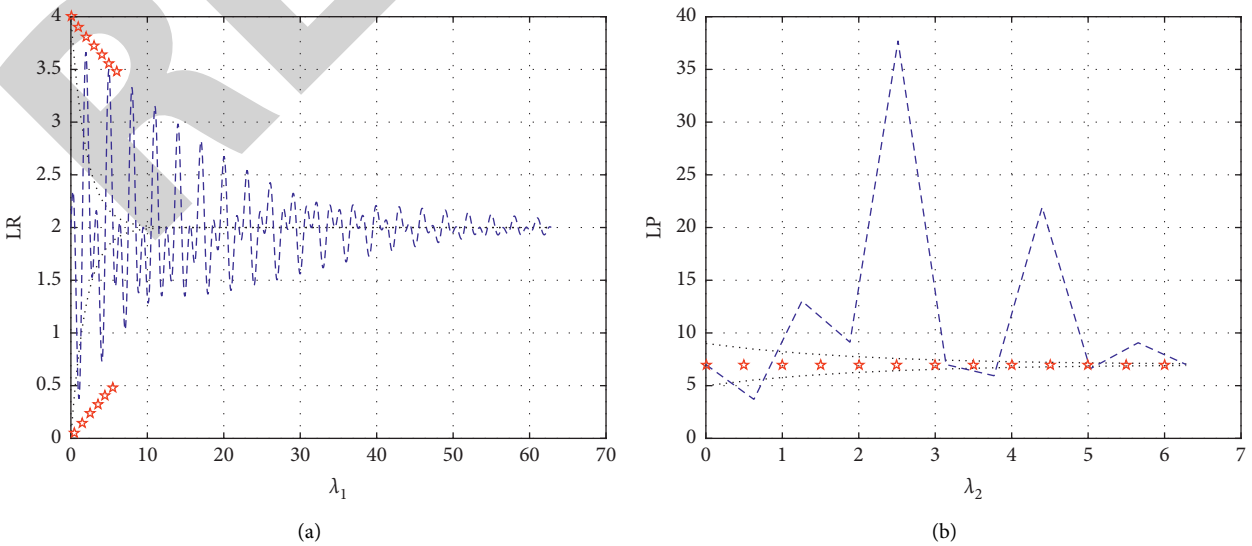

Figure 8: Variation of different λ_1, λ_2 values on the results: (a) the left is LR and (b) the right is the LP change curve.

5. Conclusion

This paper synthesizes the research results of pattern recognition and uses 1000 voices of the voice database to train the English voice emotional interaction model in virtual scenarios. Through multiple sets of comparative experiments, the feasibility of the recognition system is verified, and the noise is tested for antijamming capability and robustness. The research work of virtual scene construction adopts the currently popular three-dimensional modeling and three-dimensional virtual reality. Through experiments, the ability of emotional virtual people to recognize and feedback English speech emotion signals in virtual situations is verified. Emotional computing is a highly integrated field of research and technology. By combining computational science, psychology, and cognitive science, we will study the emotional characteristics of human-computer interaction and design a human-computer interaction environment with emotional feedback. It is possible to realize human-machine emotion. With the updating of educational teaching concepts and the advancement of science and technology, there have also been changes in the method of presenting situations. The use of various technical means to present virtual teaching situations has attracted more and more attention from the education community. Therefore, English speech emotional interaction design based on virtual situations will also receive more and more attention.

Data Availability

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

Conflicts of Interest

The authors declare that they have no conflicts of interest.

References

[1] Y. Li, X. Lu, C. Yu, H. Guo, and D. Zhang, "Research and application of the virtual simulation system teaching method in NC machining course," *International Journal of Modeling, Simulation and Scientific Computing*, vol. 9, no. 1, Article ID 1850007, 2018.

[2] N. Hasan, H. Richard, K. Manaf, and B. Fernado, "Basic skin surgery interactive simulation: system description and randomised educational trial," *Advances in Simulation*, vol. 3, no. 1, pp. 14–22, 2018.

[3] Z. J. Lei, J. J. Huang, Z. Li, L. Wang, J. Cui, and Z. Tang, "Research on collaborative technology in distributed virtual reality system," *Journal of Physics Conference*, vol. 4, no. 2, pp. 960–978, 2018.

[4] L. Huang, Y.-H. Hou, and D.-J. Zhang, "Research progress on and prospects for virtual brush modeling in digital calligraphy and painting," *Frontiers of Information Technology & Electronic Engineering*, vol. 20, no. 10, pp. 1307–1321, 2019.

[5] O. Ha and N. Fang, "Effects of interactive computer simulation and animation (CSA) on student learning: a case study involving energy, impulse, and momentum in rigid-body engineering dynamics," *Computer Applications in Engineering Education*, vol. 26, no. 5, pp. 1804–1812, 2018.

[6] P. K. Ng and B. Tung, "The importance of reward and recognition system in the leadership of virtual project teams: a qualitative research for the financial services sector," *Journal of Transnational Management*, vol. 4, no. 6, pp. 198–214, 2018.

[7] L. Qin, C. T. Wang, and C. Yao, "Research on application of location technology in 3d virtual environment modelling system for substation switch indicator," *Intelligent Automation and Soft Computing*, vol. 24, no. 1, pp. 115–122, 2018.

[8] A. R. Sinensis, H. Firman, and M. Muslim, "Reconstruction of collaborative problem solving based learning in thermodynamics with the aid of interactive simulation and derivative games," *Journal of Physics: Conference Series*, vol. 1157, pp. 032042–032054, 2019.

[9] A.-P. Correia, N. Koehler, and G. Phye, "The application of PhET simulation to teach gas behavior on the submicroscopic level: secondary school students' perceptions," *Research in Science & Technological Education*, vol. 37, no. 2, pp. 193–217, 2019.

[10] W. V. Bo, G. W. Fulmer, C. K.-E. Lee, and V. D.-T. Chen, "How do secondary science teachers perceive the use of interactive simulations? the affordance in Singapore context," *Journal of Science Education and Technology*, vol. 27, no. 6, pp. 550–565, 2018.

[11] C. Pratik, A. Akshit, and D. Varun, "Learning in an interactive simulation tool against landslide risks: the role of strength and availability of experiential feedback," *Natural Hazards & Earth System Sciences*, vol. 18, no. 6, pp. 1599–1616, 2018.

[12] S. Beloufa, J. F. Cauchard, A. KemenyVailleau, F. Mérienne, and J.-M. Boucheix, "Learning eco-driving behaviour in a driving simulator: contribution of instructional videos and interactive guidance system," *Transportation Research Part F: Traffic Psychology and Behaviour*, vol. 61, no. 2, pp. 201–216, 2019.

[13] Y. E. Papelis and M. D. Petty, "Recognizing the contributions of reviewers in publishing and peer review," *Simulation*, vol. 94, no. 4, pp. 277-278, 2018.

[14] H. Cheng, X. D. Wu, and X. Fan, "Modeling and simulation of sheet-metal part deformation in virtual assembly," *Journal of Ambient Intelligence and Humanized Computing*, vol. 10, no. 3, pp. 1231–1240, 2019.

[15] Y. Sang, W. X. Wang, and W. Sun, "Research on the development of an interactive three coordinate measuring machine simulation platform," *Computer Applications in Engineering Education*, vol. 26, no. 5, pp. 1173–1185, 2018.

[16] X. Xu, Z. Li, L. Wang, and S. Yao, "Interactive visual reality of the offshore hoisting operation and numerical modeling," *International Journal of Pattern Recognition & Artificial Intelligence*, vol. 32, no. 8, p. 1855012, 2018.

[17] L. Jiangshan and C. Ming, "Research and application of virtual simulation technology in the aerospace bearing design and manufacture," *MATEC Web of Conferences*, vol. 151, no. 8, pp. 04002–04014, 2018.

[18] L. Yunyue and Y. Fang, "Research and application of paper mill simulation systems based on virtual DPU technology," *Agro Food Industry Hi Tech*, vol. 28, no. 1, pp. 429–432, 2017.

[19] A. Tilbrook, K. T. Dwyer, and J. A. Parson, "A review of the literature-the use of interactive puppet simulation in nursing education and children's healthcare," *Nurse Education in Practice*, vol. 22, pp. 73–79, 2017.

[20] A. Kageyama, Y. Tamura, and T. Sato, "Scientific visualization in physics research by CompleXcope CAVE system," *Transactions of the Virtual Reality Society of Japan*, vol. 4, no. 4, pp. 717–722, 2017.

[21] Y. Peng, Y. Y. Ma, and J. Shan, "The application of interactive dynamic virtual surgical simulation visualization method," *Multimedia Tools and Applications*, vol. 76, no. 23, pp. 25197–25214, 2017.

[22] B. U. Seeber and S. W. Clapp, "Interactive simulation and free-field auralization of acoustic space with the rtSOFE," *The Journal of the Acoustical Society of America*, vol. 141, no. 5, p. 3974, 2017.

[23] D. Wang, "Use contexts and usage patterns of interactive case simulation tools by HIV healthcare providers in a statewide online clinical education program," *Studies in Health Technology & Informatics*, vol. 245, pp. 1242–1256, 2017.

[24] J. Chu, L. Gao, Y. Niu, and G. Li, "Research and application of virtual measuring instrument based on BPNN algorithm in starch industry," *C e Ca*, vol. 42, no. 2, pp. 512–515, 2017.

[25] B. Qu, Z. Zhang, F. Li et al., "Research and application of oil operator training system based on virtual reality technology," *Journal of Petrochemical Universities*, vol. 30, no. 1, pp. 54–59, 2017.

[26] M. Y. Kim, Y. Lee, and D. Lee, "Haptic rendering and interactive simulation using passive midpoint integration," *The International Journal of Robotics Research*, vol. 36, no. 12, pp. 1341–1362, 2017.

[27] Y. S. Lee, H. J. Yap, and R. Singh, "Implementation of a voice-control system for issuing commands in a virtual manufacturing simulation process," *Advanced Materials Research*, vol. 980, pp. 165–171, 2014.

[28] K. Mcmanus, N. R. Mitchell, and N. Tracy-Ventura, "Understanding insertion and integration in a study abroad context: the case of english-speaking sojourners in France," *Revue française De linguistique Appliquée*, vol. XIX, no. 2, pp. 97–116, 2014.