Hindawi
Advances in Multimedia
Volume 2022, Article ID 5603992, 8 pages
https://doi.org/10.1155/2022/5603992

Research Article

Computer Speech Recognition Technology and Graphic Shape Design

Yuxi Niu

Eurasia University in Xi'an, Xi'an, Shaanxi 710000, China

Correspondence should be addressed to Yuxi Niu; 201904217320@stu.zjsru.edu.cn

Received 22 February 2022; Revised 12 April 2022; Accepted 19 April 2022; Published 11 May 2022

Academic Editor: Qiangyi Li

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

In order to solve the problem of lack of multimodal emotional database, a computer speech recognition technology and graphic form design research were proposed. Using the hidden Markov model (HMM) to recognize speech emotion is mainly to provide emotion type for the subsequent expression fusion. When the emotion category in the speech is obtained, the facial animation parameter (FAP) corresponding to the speech emotion can be combined with the lip movement FAP based on the Moving Pictures Experts Group (MPEG-4) face animation standard to obtain a comprehensive FAP. The results show that the recognition effect of speech emotion recognition in the method used in this paper is relatively better than that obtained by other literature methods, and the average recognition rate reaches 70.24%, which is higher than the other three methods. It is verified that this method can present a better recognition effect.

1. Introduction

Computer speech recognition technology is the abbreviation of automatic speech recognition by machine. Speech recognition technology is related to multidisciplinary research fields. Research results in different fields have contributed to the development of speech recognition [1]. To some extent, the difficulty of making the machine recognize speech is like a person with poor foreign language listening to foreigners. It is related to the speaker, speaking speed, speaking content, and environmental conditions. The characteristics of speech signal make speech recognition difficult. These characteristics include variability, dynamics, instantaneous, and continuity [2]. As an acoustic expression, speech has become the research object of people's attention by expressing itself through direct ideological and emotional communication. Among them, the traditional signal processing theory based on linear and short-time stationary characteristics has also become an important knowledge basis for the study of speech recognition and has become a research hotspot [3].

With the continuous development of information technology, the emergence of human-computer interaction technology provides a faster way for human production and life. In the field of emotion recognition of human-computer interaction, through the use of computers to process and analyze various emotional signals, people's emotional states can be recognized so as to realize good human-computer interaction [4]. The fusion and recognition of speech recognition technology and facial expression information in graphic shape design has also become a research hotspot [5]. As the most important external feature, human face can convey a lot of nonverbal information to enhance understanding or express emotion. However, due to too many facial parameters, it becomes very cumbersome to directly manually adjust the control parameters to generate real and moving facial expressions. Voice-driven animation technology greatly reduces this workload. It directly uses speech recognition signal to drive facial animation by establishing the mapping relationship between lip movement and speech. At present, most speech-driven animations do not take into account the emotional information contained in speech recognition signal [6, 7]. Emotion is an important information resource. For the same sentence, people will show different facial expressions due to emotional joy or sadness. Computer speech recognition technology is shown in Figure 1. The use of various CAD software for auxiliary design has become a need and

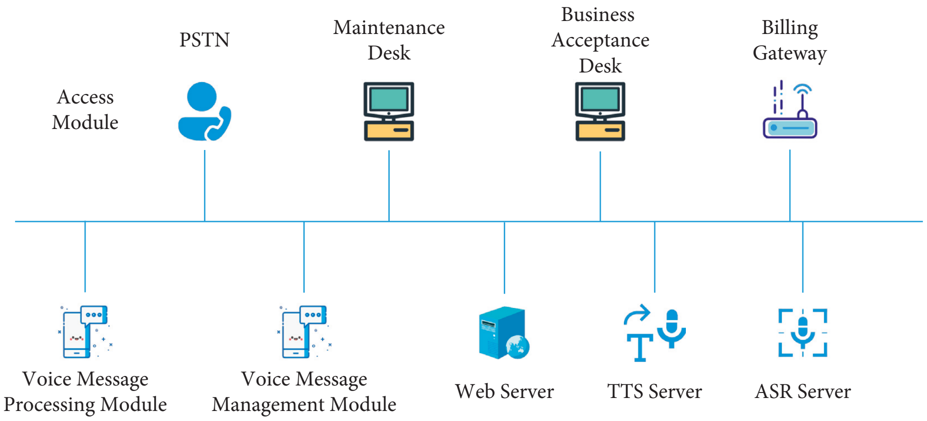

Figure 1: Computer speech recognition technology.

trend today. The human-computer interaction mode of CAD is developing in a more natural and intelligent direction. With the continuous progress and innovation of computer technology and artificial intelligence technology, CAD software is also facing a big opportunity and challenge, and a new technological revolution is emerging quietly. Scholars and engineering technicians have conducted numerous engineering practices and theoretical researches on the health monitoring of large-span steel structures but are less involved in the structural health monitoring of the whole life cycle. The application of BIM technology in the field of structural health monitoring is currently a research hotspot, but the application in the field of large-span steel structures is still lacking. Therefore, based on the characteristics of large-span space steel structure and the advantages of BIM technology in information integration and three-dimensional visualization, this paper develops a large-span space steel structure health monitoring system based on the BIM platform so as to realize the timely and accurate monitoring of large-span space steel structure monitoring data.

2. Literature Review

The tone, intensity, amplitude, fundamental frequency, resonance peak, and duration of people in different emotional states are different. Through the analysis and calculation of these structural characteristics, the emotional content contained in speech can be recognized. At present, there are many researches on speech emotion recognition. Ozseven took the pitch frequency related information as the main feature parameter, used three classifiers, maximum likelihood Bayesian classification, kernel regression, and k-nearest neighbor, to identify four emotions, fear, anger, sadness, and happiness, and achieved a recognition rate of 60% ~ 65% [8]. Poorna and Nair believe that speech, as an acoustic expression, has become the research object of people's attention through direct ideological and emotional communication [9]. Nordström and Laukka extracted effective information in the nonlinear analysis of Chinese speech through the research and calculation of correlation dimension [10]. Jermsittiparsert et al. quantitatively described the chaotic characteristics of speech signals through the fractal dimension of a single dimension [11]. Shaqra et al. introduced the nonlinear characteristics of speech signal multifractal into the research of emotion recognition and achieved a good recognition effect [12]. Konduru and Mazher Iqbal proposed an automatic speech emotion recognition system based on phase feature. The experiment shows that, compared with the ordinary system using only amplitude information, the speech emotion recognition system combining phase information and amplitude information can significantly improve the recognition performance [13]. Ahalawat and Mondal used quadratic feature selection to comprehensively select feature subsets and combined it with kernel fusion for speech emotion recognition. They carried out experiments in the emotional speech database, which not only effectively improved the accuracy of emotion recognition but also had certain robustness to speech noise [14]. Selvaraju et al. proposed a feature extraction method based on a pyramid structure double sparse learning model for speech emotion recognition. The processing task of this method is the array county with right extrusion [15]. Rosa Velardo and Frutos-Escrig used the nonlinear features of speech signals to recognize four kinds of emotional speech. Through the fusion of different features, an average recognition rate of 87.62% can be obtained. After so many years of development, speech recognition technology has become more and more mature and has made great achievements in various fields such as emotion model, emotion database, and emotion characteristics [16]. Buonamici et al. believe that the quality of speech sample data recorded in speech emotion database directly affects the effect of speech emotion recognition. Therefore, how to select or establish an appropriate emotional speech database must also be paid attention to [17]. Based on this research, this paper proposes a research on computer speech recognition technology and graphic shape design using hidden Markov model (HMM) for speech emotion recognition. Because the research purpose is for mobile RBT entertainment, emotions are divided into five categories, anger, joy, loveliness, helplessness, and excitement, and a set of 66-dimensional facial animation parameters (FAP) based on Moving Pictures Experts Group (MPEG-4) standard is established for each emotion to generate the corresponding expression animation and realize the direct mapping of voice emotion and face animation parameters. After a segment of speech is given, the expression FAP can be obtained from speech emotion recognition, and the lip movement FAP can be obtained from the speech recognition part of the system. Then, the multisource information of facial expression can be synthesized according to the comprehensive expression function obtained from the speech signal. Finally, the comprehensive FAP can be obtained to drive the two-dimensional face grid to generate realistic facial animation.

3. Research Methods

3.1. System Introduction. Figure 2 shows the general diagram of the face animation system. The system includes two parts of speech and two-dimensional face picture processing. The processing of speech part is divided into three parts as follows: (1) Speech emotion recognition. This paper selects about 180 RBT voices of different speakers and divides them into five categories according to their emotions: anger, joy, loveliness, helplessness, and excitement. Because the difference between joy, loveliness, and excitement is not particularly obvious, they all belong to joy first. Therefore, firstly, an HMM is trained to recognize anger, joy, and helplessness [18]. In order to meet the needs in the application of RBT animation, we further extract more advanced emotional feature parameters for three categories of emotions, joy, loveliness, and excitement, and train HMM recognition for each category. For an input speech, the six HMM can recognize its specific emotion category and get the corresponding expression FAP. (2) Speech recognition. After syllable division and syllable recognition, lip FAP synchronized with RBT speech signal can be obtained. (3) Comprehensive expression function. After analyzing the strength of RBT voice, the comprehensive expression function can be obtained. From this function, the total FAP can be obtained by fusing expression FAP and lip movement FAP [19].

For two-dimensional face images, 118 feature points are extracted by the face detection method to form a face grid. Finally, the face grid is driven by the total FAP to obtain the final face animation.

The main innovation points of this paper are summarized as follows (see formula (1)):

$$\min_{\theta} \frac{1}{2}\sum_{i=1}^{n}\left\| f_{\theta}(x_i) - y_i \right\|_F^2. \tag{1}$$

Finally, the optimal parameters θ and α are found by solving the following optimization problems (see formula (2)):

$$\min_{\theta,\alpha} \frac{1}{2}\sum_{i=1}^{n}\left\| f_{\theta}(\mu_{\alpha}(x_i)) - y_i \right\|_F^2. \tag{2}$$

In order to simulate the influence of noise in the real environment, this paper uses a variety of noise-adding intervention methods. After the noise is added, the speech signal can be described as shown in (3) and (4):

$$\mathrm{Dist}(x) = p_i \cdot T(x). \tag{3}$$

$$T \in \{\mathrm{Tem}(\cdot), \mathrm{Rever}(\cdot), \mathrm{Fre}(\cdot), \mathrm{Add}(\cdot)\}, x \in R. \tag{4}$$

At the same time, the operation speed of the convolutional network is greatly improved. At the same time, in order to increase the depth of the network and reduce the information loss and gradient disappearance and gradient explosion, a residual structure is introduced to build a residual gated convolutional network, and its mathematical expression is shown in (5) and (6):

$$h_i(X) = (X * W + b) \otimes \sigma(X * V + c). \tag{5}$$

$$h_{i+1}(X) = h_i(X) + h_{\mu}(X). \tag{6}$$

The mathematical expression of the basic structure of the recurrent network based on LSTM is shown in formulas (7), (8), and (9):

$$i_i = \varphi(U_i[h_{i-1}, x_i] + b_i). \tag{7}$$

$$f_i = \varphi\left(U_i\left[h_{f-1}, x_i\right] + b_f\right). \tag{8}$$

$$o_i = \varphi(U_o[h_{i-1}, x_i] + b_o). \tag{9}$$

Among them, φ is the activation function, and x_i represents the vector composed of the second dimension and the third dimension in the input matrix X. U_i represents the weight matrix of the input gate; U_f represents the weight matrix of the forget gate; U_O represents the weight matrix of the output gates b_i, b_f, and b_o, which represent the bias weight.

3.2. Speech Emotion Recognition Technology. Speech emotion recognition is a relatively new research field. It mainly analyzes the change law of speech corresponding to emotion, uses a computer to accurately extract emotion features from speech, and determines the emotion category of the tested object according to these features. The speech emotion recognition studied in this paper is mainly for mobile phone RBT, establishes the direct mapping relationship between RBT speech emotion and facial animation parameters, and provides expression categories for subsequent expression fusion [20].

3.2.1. Emotional Speech Database. The establishment of speech emotion database is the basis of speech emotion recognition [21]. According to the RBT speech samples, it is divided into five kinds of emotions: anger, joy, loveliness, helplessness, and excitement. Cool Edit Pro 2.1 is used to edit about 300 emotional voices that belong to the above five types of emotions from the RBT voice and save them as WAV files (16000 Hz, 16 bits, mono). The selection of voice clips mainly follows the following two principles: (1) The sentences cannot have clear semantic tendency. Only in this way can we ensure that the construction of voice database will not affect the judgment of experimenters. (2) The time of each voice is controlled within 7 s, which is not conducive to the expression of emotion and will weaken the characteristic

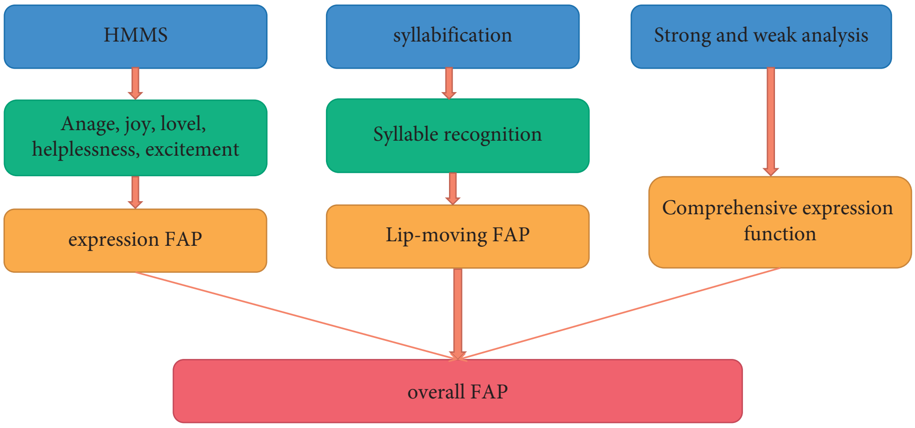

Figure 2: General diagram of system.

parameters used for emotion judgment. In order to test the effectiveness of the collected emotional speech, listening experiments are carried out in this paper. Three experimenters were invited to listen to these emotional sounds, and they were asked to say the emotional categories of the played sounds through subjective judgment [22, 23]. In addition to those sentences that are not within the five kinds of emotions or the fact that the emotion type is not obvious, they voted to divide these sounds into five kinds of emotions: anger, joy, loveliness, helplessness, and excitement. Finally, 179 sentences were selected as the emotional speech database of this paper, as shown in Table 1.

3.2.2. Emotional Feature Parameter Extraction. In the research of speech emotion recognition, the extraction of speech feature parameters plays a decisive role in the recognition effect. It can accurately reflect the speech parameters of emotional features and significantly improve the recognition rate. In this paper, four kinds of parameters such as pronunciation rate, pitch frequency, formant, and Mel frequency cepstrum are extracted as emotional features.

(1) Pronunciation Rate. Pronunciation rate belongs to the characteristics of speech signal in time structure. It reflects the urgency of people's mood when talking and will vary with different emotional states. In this paper, the pronunciation rate refers to the number of syllables contained in the speech signal per unit time. The number of syllables refers to the number of peaks in the speech signal whose peak exceeds a certain threshold. The unit time of speech signal also includes the mute part of speech because the mute part also contributes to emotion.

(2) Pitch Frequency. Pitch frequency is one of the important features reflecting emotional information. In this paper, the pitch frequency is calculated frame by frame by cepstrum method, and the pitch frequency curve is median filtered and linear smoothed. The positive and negative mean values of fundamental frequency change rate in unit time period are extracted as characteristic parameters. The fundamental frequency change rate here refers to the first-order difference of the fundamental frequency of each frame speech signal.

Table 1: Sample number of speech emotion database.

Emotion category	Anger	Delighted	Lovely	Resignation	Excitement
Sample individual	44	31	34	33	26

(3) Formant. Formant is an important parameter reflecting the characteristics of sound channel. Different emotions may cause different changes in the vocal tract and also lead to different resonance peaks of speech signals under different emotional states. In this paper, the 16th order prediction coefficient is obtained by the linear prediction method, then the frequency response curve of the channel is estimated by the prediction coefficient, and finally, the frequency and bandwidth of each formant are calculated by the peak detection method. In the classification of anger, joy, and helplessness, only the mean value of the first formant frequency per unit time is selected as the characteristic parameter. Because the mean value of the frequency and bandwidth of the first three formants has a good effect on further subdividing the three categories of speech emotion of joy, loveliness, and excitement, it is selected as the characteristic parameter when subdividing these three categories of emotional speech.

(4) Mel Frequency Cepstrum. MFCC (Mel frequency cepstral coefficients) is a cepstral characteristic parameter extracted in the Mel scale frequency domain, which describes the nonlinear characteristics of human ear's perception of frequency. Firstly, the speech signal is processed, then the window is added according to the window length of 30 ms and the window shift of 10 ms, and 12 triangular filters are selected to form a filter bank according to the application characteristics. When the signal passes through the filter bank, the weighted sum of all signal amplitudes within the frequency bandwidth of each filter is taken as the output of each band-pass filter, and then the logarithm of all filter outputs is calculated. Finally, the dimension of the feature

vector is reduced by discrete cosine transform, and the first 12-dimensional MFCC is obtained as the feature parameter.

In the first step (recognizing anger, joy, and helplessness), 16-dimensional feature vectors of the above four types of emotion parameters are extracted for each speech. When voice emotion is recognized as joy, the second step is required to identify which of the three categories of joy, loveliness, and excitement. At this time, the extracted emotion parameter is a 21-dimensional feature vector [24].

3.2.3. Realization of Speech Emotion Recognition. Considering that, in most cases, people's emotions will not change in a short time, this system recognizes each RBT speech sample as only one kind of emotion. Speech emotion recognition is essentially a process of pattern recognition. At present, there are linear discriminant classification, k-nearest neighbor method, SVM, and other methods. Based on the wide application and excellent performance of HMM in the field of speech recognition, it is selected as the recognizer in this paper.

HMM is essentially a dual stochastic process finite-state automata, including the stochastic process of state transition and the stochastic process of observation output. The stochastic process of state transition is implicit, which is expressed by the stochastic process of observation sequence. HMM can be represented by triples: $\mu = (A, B, \prod)$, where A is the state transition matrix, B is the observed output probability matrix of the state, and $\prod$ is the initial distribution probability matrix of the state.

When using HMM for speech emotion recognition, it is necessary to establish an HMM for each kind of emotion. Because left-right HMM has the advantages of low computational cost and few iterations, this paper uses five kinds of emotional speech training in speech database: anger, joy, loveliness, helplessness, and excitement to represent the left-right HMM of each emotion. Because the recognition rate of the model is related to the number of states, generally, more states can get a higher recognition rate. However, considering the efficiency and complexity, this paper selects 7 states for each model, uses a Gaussian distribution to estimate the output probability density function, inputs the extracted characteristic parameters of each emotion into each HMM at one time, and is trained by Baum Welch algorithm. Finally, the model parameters of each emotion are obtained [25].

Four-fifths of the speech database in this paper are used as the training sample, and the other one is used as the test. In order to avoid contingency, five tests are conducted respectively, corresponding to test 1 ~ test 5 in Table 2. In order to verify the effectiveness of various emotional features, experiments were carried out under five different combinations. In the first step, the recognition rate of anger, joy, and helplessness is shown in Table 2. The emotional feature vectors are 12-dimensional MFCC (12 dimensional), 12-dimensional MFCC combined with 1-dimensional formant (13 dimensional), 12-dimensional MFCC combined with 2-dimensional pitch frequency (14 dimensional), 12-dimensional MFCC combined with 1-dimensional common resonance peak and 2-dimensional pitch frequency (15 dimensional), and 12-dimensional MFCC combined with 1-dimensional formant, 2-dimensional wiki audio rate, and 1D pronunciation rate (16 dimensional) 1 from the recognition rate; when the 16D emotional feature vector is selected, the total average recognition rate of the five experiments is the highest, 94.37%, reaching a high recognition rate.

Table 2: Recognition rate in combination of different eigenvectors%.

	12D	13D	14D	15D	16D
Test 1	86.77	86.77	92.83	90.81	95.77
Test 2	92.83	90.81	83.74	77.81	80.14
Test 3	90.81	86.77	90.81	92.83	100.00
Test 4	80.71	92.83	80.71	86.77	95.77
Test 5	90.21	90.81	77.68	80.81	95.77

According to the experimental data in the first column of Table 2, when only 12-dimensional MFCC is used as the feature vector, the average recognition rate has reached 88.97%. Most of the existing speech emotion recognition only takes the statistics related to pitch frequency as the feature parameters, which can better reflect the emotional characteristics of the speech without background noise interference. However, for the RBT speech with a large number of types of background music interference in this speech library, the feature vectors obtained from the combination of MFCC, formant, pitch frequency, and speech rate are more anti-interference. It can better represent the emotional information of RBT.

When voice emotion is recognized as joy, the second step is required to identify which of the three categories of joy, loveliness, and excitement. The comparison of test results using the original 16-dimensional feature vector and 21-dimensional feature vector is shown in Table 3. When using the 21-dimensional feature vector, the recognition rate is significantly improved. At this time, the total average recognition rate of five types of emotions can reach 94.44%. After obtaining more accurate emotion types, it can be used to drive face images to generate expression animation.

Using the emotional feature vector selected in this paper, SVM is used to train and recognize the RBT speech database. Compared with the method in this paper, the results are shown in Figure 3. Although the training samples in this RBT database are not enough, the SVM requires a large number of training samples, the RBT voice itself contains background noise, and it can still be seen that the HMM in this paper can obtain a higher recognition rate than SVM. This shows that HMM is more suitable for emotion recognition than SVM for RBT speech with music background or noise interference.

3.3. Facial Animation Expression Fusion. The face animation system in this paper is based on the MPEG-4 standard. Some standards related to face are specially defined in MPEG-4. It uses 68 personal face animation parameters FAP to describe various complex facial expressions and uses face definition parameters (FDP) to describe face model and face texture. This paper classifies the five kinds of emotion in the speech database and establishes a set of corresponding expression FAP values. When the emotion categories are obtained from

Table 3: Comparison of 16D and 21D feature vector recognition rate%.

	Test 1	Test 2	Test 3	Test 4	Test 5	Average recognition rate
16D	94.33	94.20	88.78	66.56	88.78	86.51
21D	94.33	94.33	88.78	94.33	100.00	94.33

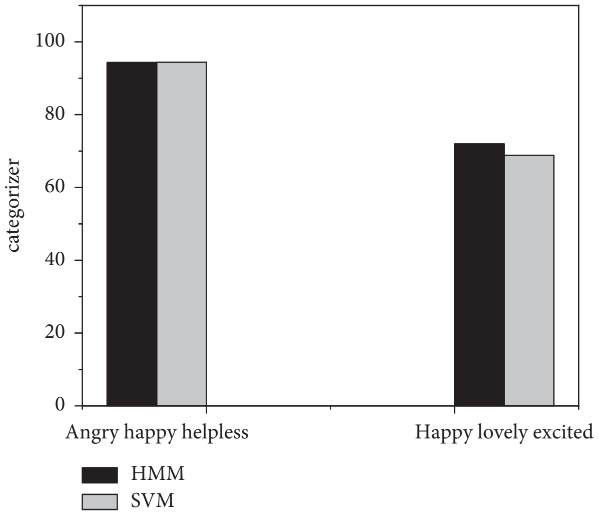

Figure 3: Comparison of recognition rate between HMM and SVM%.

the above speech emotion recognition, the corresponding expression FAP can be obtained to form a mapping relationship. Based on the current research on Chinese visual elements and MPEG-4 standard, this paper simply divides Chinese into six basic lip shapes and establishes the corresponding lip FAP value. Because the final speech-driven facial animation should include lip movement and expression at the same time, it is necessary to integrate the two. In this paper, the comprehensive expression function is used as the constraint condition for multisource information fusion of facial expression [26].

3.3.1. Comprehensive Expression Function. The comprehensive expression function mainly reflects the ups and downs of human facial expression when speaking. In general, the change of emotional intensity is consistent with the change of speech intensity, and the change of speech intensity is generally periodic and changes from weak to strong and then from strong to weak in each cycle. Therefore, the chi-square distribution function is used to describe this change in the cycle. The probability density function of the chi-square distribution is as follows:

$$f(y) = \begin{cases} \dfrac{1}{2^{n/2}\Gamma(n/2)} y^{n/2-1_{e}-y/2}, & y > 0, \\ 0, & y \leq 0. \end{cases} \tag{10}$$

The function is shown in Figure 4, where the y-axis is the time axis, and the probability density function curve of chi-square distribution when $n = 5$ conforms to this change trend.

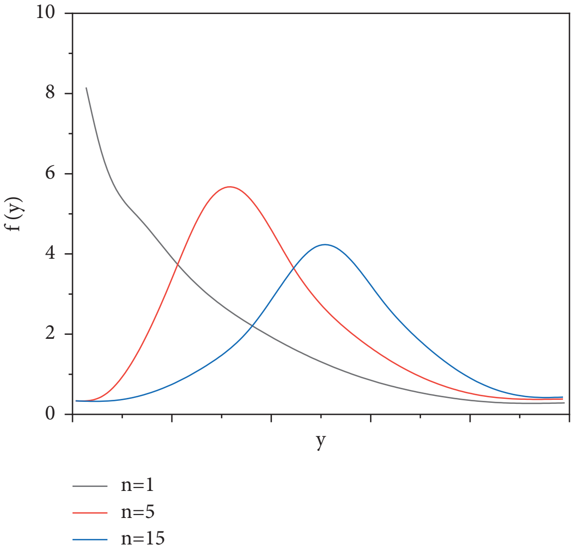

Figure 4: Probability density of chi-square distribution.

The relevant information of the speech cycle is obtained by analyzing the strength change of specific speech. Firstly, the RBT voice signal is sampled to obtain the sound intensity at each sampling point, and the peak of the voice is obtained based on the sound intensity at each sampling point so that each peak corresponds to a syllable in the voice and the peak amplitude corresponds to the pronunciation intensity of this syllable. Secondly, set the threshold of wave peak and wave trough, take the place where the wave peak in speech is greater than the threshold as the peak point of chi-square distribution in this cycle, its value is the ratio of the wave peak to the maximum peak value, and search from left to right where the wave peak is less than the valley threshold as the beginning and end of this cycle. By inserting the corresponding chi-square distribution function in each cycle, the comprehensive expression function of a speech can be obtained.

4. Result Discussion

For the bimodal emotion recognition composed of speech and facial expression, after extracting the speech features and facial expression image features, the features of the two modes are fused based on the feature layer, and then the six emotions in the database are classified and recognized by random forest recognition classifier. In order to ensure the accuracy of the recognition results and prevent the phenomenon of overfitting, ten cross-validation methods are used in the experiment. Different feature fusion algorithms are used to compare the emotion recognition effects of speech, face, and expression bimodal emotion recognition. Through the above methods, we can get the comparison of the recognition effects of six emotions: single-mode emotion recognition of speech and facial expression and decision-making level fusion of two-mode emotion recognition through quadrature rules. The extraction method in this paper is compared with other methods, and the comparison is shown in Table 4.

Table 4: Comparison of accuracy between this method and other methods.

	Speech recognition	Face recognition expression	Fusion of speech recognition and face recognition
Clustering and machine learning methods	34.00	24.00	66.00
Cartoon face animation system	32.00	36.00	70.00
Animation system	61.80	53.60	70.20
Paper method	70.24	62.44	77.51

From the table, in the same database, the recognition effect of speech emotion recognition in the method used in this paper is relatively better than that obtained by other methods, and the average recognition rate is 70.24%, higher than the other three methods. At the same time, facial expression recognition using this method has a better recognition effect than other facial expression recognition methods, and the average recognition effect is 62.44%. Finally, compared with other emotion recognition effects, the emotion recognition combining speech and facial expression can get a better recognition effect, and the average recognition rate is 77.51%.

5. Conclusion

Realistic computer face animation is one of the most basic problems in the field of computer graphics, and it is widely used in human-computer interaction, virtual reality, and other fields. This paper implements a method and system for generating face animation driven by computer speech recognition technology and emotion. HMM is selected as the classifier and trained to recognize five kinds of emotions: anger, joy, loveliness, helplessness, and excitement in the speech database, and a set of corresponding expression facial animation parameters (FAP) is established for each kind of emotion. The speech strength is analyzed to obtain the comprehensive expression function, this function is used to fuse the expression FAP and lip movement FAP to realize the multisource information synthesis of facial expression, and the comprehensive FAP is obtained to drive the facial mesh to generate animation. The experimental results show that the face animation generated by the system also has high realism. There is an obvious nonlinear relationship between the temperature and the strain change value of the large-span space steel structure. The neural network model can better fit the nonlinear relationship between the temperature and the strain and can be used to predict the temperature effect of the structure. The nonlinear relationship roughly presents a spindle shape, and the smaller the spindle shape area is, the more it tends to a straight line, indicating that the linear relationship between temperature and strain changes is stronger, and vice versa, the nonlinear relationship is stronger. The face animation in this paper can also further consider facial skin folds, adding tooth models, and so on.

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 no conflicts of interest.

Acknowledgments

This study was funded by the Key Course Construction Project of Eurasia University in Xi'an in 2019: Design Foundation (Project no. 2019KC015).

References

[1] J. Jiang and H. H. Wang, "Application intelligent search and recommendation system based on speech recognition technology," *International Journal of Speech Technology*, vol. 24, no. 1, pp. 23–30, 2021.

[2] I. Calvo, P. Tropea, M. Viganò, M. Scialla, and M. Corbo, "Evaluation of an automatic speech recognition platform for dysarthric speech," *Folia Phoniatrica et Logopaedica*, vol. 73, no. 5, pp. 1–10, 2020.

[3] T. Zia and U. Zahid, "Long short-term memory recurrent neural network architectures for Urdu acoustic modeling," *International Journal of Speech Technology*, vol. 22, no. 1, pp. 21–30, 2019.

[4] L. Wei, "Study on the application of cloud computing and speech recognition technology in English teaching," *Cluster Computing*, vol. 22, no. 4, pp. 9241–9249, 2019.

[5] X. Sun, T. Hong, C. Li, and F. Ren, "Hybrid spatiotemporal models for sentiment classification via galvanic skin response," *Neurocomputing*, vol. 358, pp. 385–400, 2019.

[6] R. Y. Rumagit, G. Alexander, and I. F. Saputra, "Model comparison in speech emotion recognition for Indonesian language," *Procedia Computer Science*, vol. 179, no. 1, pp. 789–797, 2021.

[7] M. Gomathy, "Optimal feature selection for speech emotion recognition using enhanced cat swarm optimization algorithm," *International Journal of Speech Technology*, vol. 24, no. 1, pp. 155–163, 2021.

[8] T. Özseven, "A novel feature selection method for speech emotion recognition," *Applied Acoustics*, vol. 146, pp. 320–326, 2019.

[9] S. S. Poorna and G. J. Nair, "Multistage classification scheme to enhance speech emotion recognition," *International Journal of Speech Technology*, vol. 22, no. 2, pp. 327–340, 2019.

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

[11] K. Jermsittiparsert, A. Abdurrahman, P. Siriattakul, L. A. Sundeeva, and A. Maseleno, "Pattern recognition and features selection for speech emotion recognition model using deep learning," *International Journal of Speech Technology*, vol. 23, no. 4, pp. 1–8, 2020.

[12] F. A. Shaqra, R. Duwairi, and M. Al-Ayyoub, "Recognizing emotion from speech based on age and gender using

hierarchical models," *Procedia Computer Science*, vol. 151, no. C, pp. 37–44, 2019.

[13] A. K. Konduru and J. L. Mazher Iqbal, "Multidimensional feature diversity based speech signal acquisition," *International Journal of Speech Technology*, vol. 23, no. 3, pp. 527–535, 2020.

[14] N. Ahalawat and J. Mondal, "An appraisal of computer simulation approaches in elucidating biomolecular recognition pathways," *Journal of Physical Chemistry Letters*, vol. 12, no. 1, pp. 633–641, 2020.

[15] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: visual explanations from deep networks via gradient-based localization," *International Journal of Computer Vision*, vol. 128, no. 2, pp. 336–359, 2020.

[16] F. Rosa-Velardo and D. D. Frutos-Escrig, "Decidability and complexity of petri nets with unordered data," *Theoretical Computer ence*, vol. 412, no. 34, pp. 4439–4451, 2020.

[17] F. Buonamici, R. Furferi, L. Governi et al., "A practical methodology for computer-aided design of custom 3d printable casts for wrist fractures," *The Visual Computer*, vol. 36, no. 2, pp. 375–390, 2020.

[18] M. S. Iglesias, L. Trevisan, and F. X. Xhafa, "The approximability of non-boolean satisfiability problems and restricted integer programming," *Theoretical Computer Science*, vol. 332, no. 1, pp. 123–139, 2019.

[19] G. D'Agostino and G. Lenzi, "On the -calculus over transitive and finite transitive frames," *Theoretical Computer Science*, vol. 411, no. 50, pp. 4273–4290, 2019.

[20] L. Santocanale and A. Arnold, "Ambiguous classes in -calculi hierarchies," *Theoretical Computer Science*, vol. 333, no. 1, pp. 265–296, 2019.

[21] Y. Xuan, G. Han, L. Li, A. Qian, and W. Zhang, "Igrc: an improved grid-based joint routing and charging algorithm for wireless rechargeable sensor networks," *Future Generation Computer Systems*, vol. 92, pp. 837–845, 2019.

[22] G. J. Nalepa, K. Kutt, and S. Bobek, "Mobile platform for affective context-aware systems," *Future Generation Computer Systems*, vol. 92, pp. 490–503, 2019.

[23] F. H. Kamaru Zaman, "Locally lateral manifolds of normalised gabor features for face recognition," *IET Computer Vision*, vol. 14, no. 4, pp. 122–130, 2020.

[24] M. M. Al-Sayed, H. A. Hassan, and F. A. Omara, "An intelligent cloud service discovery framework," *Future Generation Computer Systems*, vol. 106, pp. 438–466, 2020.

[25] X. Qin, Y. Shi, K. Lyu, and Y. Mo, "Using a tam-toe model to explore factors of building information modelling (bim) adoption in the construction industry," *Journal of Civil Engineering and Management*, vol. 26, no. 3, pp. 259–277, 2020.

[26] S. G. Mahiwal, M. K. Bhoi, and N. Bhatt, "Evaluation of energy use intensity (eui) and energy cost of commercial building in India using bim technology," *Asian Journal of Civil Engineering*, vol. 22, no. 3, pp. 1–18, 2021.