Scalable Computing: Practice and Experience, ISSN 1895-1767, http://www.scpe.org
© 2024 SCPE. Volume 25, Issues 1, pp. 299–311, DOI 10.12694/scpe.v25i1.2296

RESEARCH ON THE APPLICATION OF SPEECH DATABASE BASED ON EMOTIONAL FEATURE EXTRACTION IN INTERNATIONAL CHINESE EDUCATION AND TEACHING

XIANGLI ZHANG*

Abstract. The advanced analysis of the relationship between acoustic and emotional characteristics of speech signals can effectively improve the interactivity and intelligence of computers. Given the current status of speech recognition and the problems encountered in international Chinese education, the study proposes to extract emotional characteristics to achieve speech construction of the database. Based on considering the emotional characteristics of speech, a hybrid algorithm based on spectral sequence context features is proposed. The DBN-BP algorithm is used to process emotional data of different dimensions, and a speech database is constructed. After testing and analyzing the algorithm model, it is found that the dynamic recognition accuracy of the DBN-BP model fused with emotional features is over 90%, and the negative emotion recognition rates in the three databases are all above 60%. At the same time, the accuracy rate of the model in the algorithm comparison experiment remains above 85%, the data information extraction is relatively complete, and the average test time of less than 1s is less than 3%. The speech database based on multi-emotional feature extraction can effectively provide a new reference for the improvement of the quality of Chinese international education and the improvement of the speech recognition system.

Key words: Emotional characteristics; Speech database; Chinese education; International teaching; DBN-BP algorithm

1. Introduction.

With the continuous development of artificial intelligence technology, speech recognition technology has become increasingly mature, and it has been applied in various fields such as medical services, remote education, transportation, etc. However, due to the complexity and differences of emotional activities, the current development of speech emotion recognition still faces significant limitations. At the same time, internal and external factors such as environmental noise interference, emotional fluctuations in the speaker, channel distortion in the speech library, native language habits, individual differences in the speaker, and language learning environment can have a significant impact on speech recognition and application effectiveness, and inevitably lead to language understanding ambiguity between the communication parties. In current international Chinese education and teaching, the communicative function of language teaching has not been given sufficient attention, and the breadth and depth of Chinese language and culture have further increased the difficulty of learning teaching aids. Murvey B scholar found that some foreign college students who went to China to study had different cognitive attitudes and life paths before and after graduation. They were often in ambivalence in the learning of cultural knowledge and language, and had poor initiative [1]. One of the key aspects of increasing the quality of international Chinese education is to make this cultural dissemination method more vivid. Building a voice database can effectively reduce teaching difficulty. The Kaur J scholar team described the automatic spectral speech recognition technology and studied the current development of tonal language [2]. Zehra W scholars introduced ensemble learning into cross corpus and multilingual emotion recognition, and proved that the mutual application of different corpus data can improve the accuracy of corpus data training [3]. Emotional features refer to a series of features in human speech that can reflect emotional states, including tone, volume, speed, intonation, etc. These features can identify the speaker's emotional state in speech, such as anger, joy, sadness, etc. And a speech database refers to a resource library that collects and stores a large amount of speech data, which contains pronunciation samples from different populations. Analyzing different speech samples in the speech database can extract features related to emotions, allowing computers to automatically recognize the speaker's emotional state. There are differences in acoustic feature patterns under different emotional states. Research suggests that there is a close relationship between emotional

*School of Chinese Language and Literature, Panzhihua University, Panzhihua 617000, China (xianglidream@sina.com)

features and speech databases. Therefore, to further improve the quality and effectiveness of Chinese education and teaching, the study starts from the dimension of emotional feature extraction to achieve the construction of speech databases.

2. Literature review. The emotional feature is a kind of information feature with diversity and complexity. Most scholars have studied its feature recognition. Among them, Fan X and other scholars have used the singular value non-solution algorithm after the subband division of the speech signal with the help of wavelet packet change. The dictionary training was performed on two different feature sets, real and false. The results denoted that the classification module integrating cepstral features and sparse decomposition expressed a recognition rate of more than 75% during the experiment, which further improved the ability of the Chinese deception detection system [4]. The machine team of Pan L scholars used a machine learning algorithm to construct a model of the English part of speech and eliminates the ambiguity in the recognition based on relevant rules and phrase structure. The experimental outcomes indicated that the classification algorithm had a good application effect [5]. Scholars such as Mnassri A used a genetic algorithm to optimize the parameters of the support vector machine, thereby improving the accuracy of speech recognition. The test findings indicated that the selected Arabic words could be effectively input after cepstrum processing, and could be used in the comparison. A high recognition rate could be achieved in a short training time [6]. The RNN transducer greatly simplified the automatic speech recognition system, but the realization of its training process was quite difficult. Based on this, scholars such as Wang S have used the learning rate decay strategy and added convolutional layers to improve the ability to understand Chinese [7]. Koduru A believed that paying attention to the extraction of speech signals could effectively understand its speech emotions. MFCC coefficients, zero-crossing algorithms, and global features were used to achieve feature extraction and information screening. Simulation findings expressed that the extraction algorithm could effectively improve happiness, etc. The general sentiment was extracted [8].

Scholars such as Kumaran U proposed to use deep C-RNN to realize the change of emotions in the classification stage, that is, to distinguish emotional features of different natures through the extraction of high-level spectral features and the learning of contextual features. The data loss value was small, and the accuracy rate of emotion recognition of speech signals was more than 80% [9]. Kerkeni L and his research team used the empirical pattern decomposition system to realize speech emotion recognition, that is, the signal was decomposed into feature modulation and emotion recognition to improve its classification performance. And the research outcomes illustrated that the system was supported by machine algorithms, in database verification. It had a recognition rate of more than 85% [10]. Aiming at the problem that the prosody and sound quality features were greatly affected by the signal-to-noise ratio (SNR) in the speech emotion recognition, Huang Y and other scholars proposed to use weighted ideas and deep belief networks to realize feature processing and fusion operations. The results showed that the feature learning structure could better reduce the interference problem of noisy environments and improve the accuracy and application performance of emotion recognition [11].

At the same time, Daneshfar F and other scholars used the QPSO algorithm to perform dimension reduction projection processing on the extracted high-dimensional rich features and improve the algorithm considering the classifier parameters. The research outcomes proved that the accuracy of the research system in the emotional speech database was better than other comparison algorithms [12]. Chen M and other scholars proposed a three-dimensional attention convolutional recurrent neural network to distinguish SER features, which reduced the interference of other irrelevant factors based on retaining information and emotional features. The experimental findings indicated that the method had high application effectiveness, and the recall rate was high [14]. Scholars such as Kwon S proposed that the deep convolutional network of INCA performed the most characteristic prediction, and collected and processed the data from the extraction of spectral and spatial domain features. The experiment outcomes expressed that the prediction system under the classifier performed more than 80% recognition rate, with good application effectiveness [15]. There were many types of research on feature algorithms for emotion recognition, but they were rarely applied in international Chinese teaching. Scholars such as Widodo HP believed that under the current globalization trend, Chinese teachers should pay attention to the construction and negotiation of their professional identity [13]. Yuan R and other scholars analyzed the cognition of college students' international courses with a new perspective of identity. The experimental findings illustrated that participants' positioning of themselves often fell into the paradox of personal roles and social roles [17]. Xinhan N scholar started with the research on student management teaching systems, constructed an intelligent analysis system based on neural network technology and emotion feature recognition algorithms, and designed a relevant scale evaluation index system using machine learning methods. The results denoted that the proposed model had good classroom application effectiveness [18]. Hu Jingchao, a scholar, combined deep learning with HMM feature algorithms to design a teaching state detection system, and completed the construction of recognition models through the collection and processing of subjective evaluation data and feature discretization. The outcomes expressed that this algorithm could effectively recognize student state features, with a recognition accuracy of over 90% [19]. Byun S W scholars used recursive neural network models to extract emotional recognition features and classify emotions from different aspects of acoustic features. The findings indicated that the accuracy of the designed system exceeded 85%, and its applicability was good [20]. Shah V scholars believed that introducing machine learning algorithms into text data analysis could effectively identify emotional states contained in information data [21].

3. Research on the construction of a speech database based on emotional features.

3.1. Extraction algorithm based on spectral sequence context mixed features. Feature extraction is an important step in speech emotion recognition and database establishment. The acoustic parameters contained in the speech signal are the main distinguishing points of different speech features. Generally, the validity and difference of the feature set are processed with generation and evaluation modules. The emotional acoustic features of speech are less likely to fluctuate due to differences in expression methods, which are largely related to the emotional attitude and emotional fluctuations presented by the speech. When different people express the same language meaning, they may unconsciously reveal individual emotional tendencies due to their own language habits and personal preferences, which are reflected in speech acoustic features such as time-frequency domain and cepstrum features. Time domain features refer to the features exhibited by speech signals within a certain time range after windowing processing. When the time domain waveform of a single frame signal crosses the time axis and causes different changes in adjacent sampling values, the speech signal exhibits high and low-frequency features. The number of changes is positively correlated with the frequency. The common time-domain features include short-time energy, average amplitude, autocorrelation, and so on. It is difficult to estimate the period of short-time autocorrelation due to large amount of calculation and long time consumption, and it is difficult to determine the appropriate size of the window length. Therefore, the study uses the short-term average amplitude difference function to calculate the period, and the calculation formula is shown in equation 3.1 [22].

$$F_n(k) = \sum_{m=N-1-k}^{N-1} |x_n(m) - x_n(m+k)| \tag{3.1}$$

In equation 3.1, $x_n(m)$ is the voice signal; $x_n(m+k)$ is the maximum delay point; N means the time; m represents frame shift. The frequency domain feature reflects the eployment of signal energy in different frequency bands, and can reflect the overall periodic performance of the signal. Part of the formula is shown in equation 3.2.

$$\begin{gathered} S_f = \textstyle\sum_{n-1}^{N} (A_i(n) - A_{i-1}(n))^2 \\ \textstyle\sum_{n-1}^{S_r} A(n) = \frac{17}{20} \sum_{n-1}^{N} A(n) \end{gathered} \tag{3.2}$$

In equation 3.2, S_f, S_r is the spectrum transition parameter and the spectrum cutoff parameter; $(A_i(n) - A_{i-1}(n))$ denotes the current amplitude spectrum of the frame number and the previous amplitude spectrum; n is the number of spectral lines [23].

$$C(n) = \mathcal{F}^{-1} \left(\ln \left|\mathcal{F}\left(x_n(m)\right)\right|\right) \tag{3.3}$$

Equation 3.3 is the cepstrum characteristic parameter, in which $F(), F^{-1()}$ respectively represent the forward and inverse changes of Fourier, and $|\mathcal{F}\left(x_n(m)\right)|$ denotes the real part of the complex number [24]. Speech signals are often continuous and whole, and the intonation and emotion between the previous frame and the

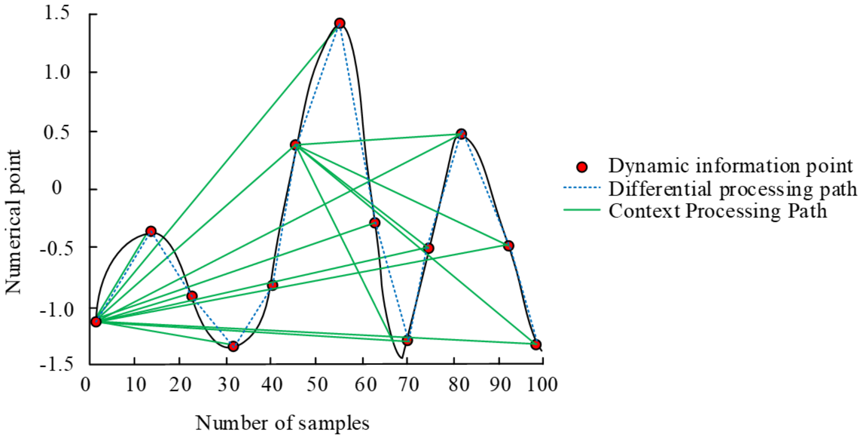


Fig. 3.1: Dynamic information processing mode

next frame are mutually influenced and run through the whole speech sequence. Therefore, the study proposes a feature extraction based on the spectrum sequence context feature (SSC). Algorithms are utilized to strengthen the grasp of dynamic correlation information between all frames. Figure 3.1 shows two information processing modes.

The contextual processing in Figure 3.1 can get the relevant dynamic information between all the frames better, and it ensures the dynamic spectral information and reduces the loss of information compared with the traditional differential processing of time series. The difference distance between the spectral sequence frames are calculated, as shown in equation 3.4.

$$\begin{gathered} D_{pq} = c_q - c_p, p, q = 1, 2, \ldots, M \\ Q = \begin{bmatrix} 0 & D_{12} & D_{13} & \ldots & D_{1M} \\ D_{12} & 0 & D_{23} & \ldots & D_{2M} \\ D_{13} & D_{32} & 0 & \ldots & D_{3M} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ D_{1N} & D_{2M} & D_{3M} & \ldots & 0 \end{bmatrix} \end{gathered} \tag{3.4}$$

In equation 3.4, N means the frame index; p, q are the spectral frame index; c_q, c_p express the spectral sequence of the corresponding frame; $D_{i,j}$ denotes the difference value between two frames; Q refers to the order vector matrix. Then, the average value of the feature set and the distance between the feature spectrum and the average spectrum are obtained, as shown in the equation 3.5.

$$\begin{array}{ll} C_{\text{avs}} & = \sum_{i=1}^{N} \frac{c_p}{N} \\ \text{diag}_p & = C_p - C_{\text{avs}}, \quad p = 1, 2, \ldots, n \\ F_m & = \begin{cases} S_{pq} + \text{diag}_p, & \text{if } p = q \\ S_{pq}, & \text{if } p \neq q \end{cases} \end{array} \tag{3.5}$$

In equation 3.5, C_{avs} indicates the average value; *diag* expresses the distance; S stands for the difference matrix; *diag* refers to the spectral center difference; F_m means the fused feature matrix. Figure 3.2 is a schematic diagram of a spectral context feature extraction process.

In Figure 3.2, the input speech signal is firstly processed by adding windows and splitting frames. The speech signal data is collected in one segment. To ensure the batch processing of the data by the programme, it needs to be transformed into the programmed data structure according to the specified length, i.e., subframe. At the same time, the signal processing requirements for continuous conditions. If the signal is disconnected during the subframe processing, it is necessary to add windows to the subframe data to better ensure the continuity of

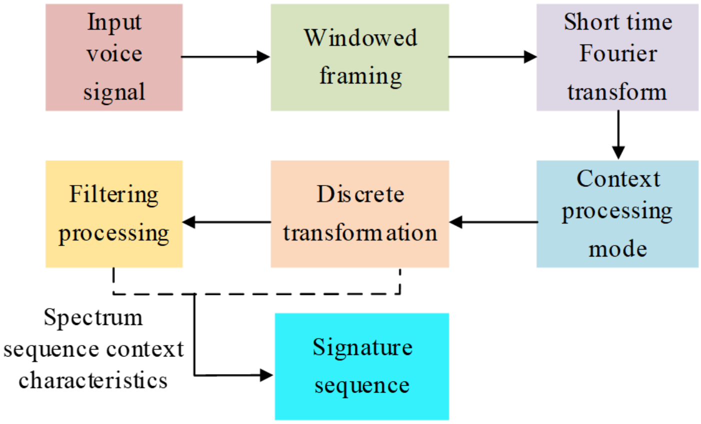


Fig. 3.2: Schematic diagram of spectrum context feature extraction process

the signal data. Subsequently, the processed speech signal data is subjected to short-time Fourier transform to better obtain the relationship between the time domain and the frequency domain. And the transformed signal is subjected to frame Gammotone filtering results and power-law compression, DCT transform, and then the set of spectral features is obtained. The set of spectral features is subjected to discrete cosine transform and context processing to obtain the context features. The different distances between the context features can be used as the basis for their differential processing, and the distance between each frame and the average value can be calculated to obtain the feature sequence needed for the study.

The rotation-invariant performance of contextual features can reduce the impact of the complexity of the vocal system and the diversity of speech content on speech recognition in complex background environments. The dimensional feature data extraction is performed on the context features by spatiotemporal Gabor filtering, that is, the temporal modulation filter is used as a row vector, and the frequency domain modulation filter is represented as a column vector, and is convolved with the feature channel and frame, respectively. The calculation method is shown as equation 3.6.

$$\mathrm{filtcrcdSSC}(g,z) = \sum_{i,j} \mathrm{SSC}(g-i, z-j) \cdot \mathrm{filtcrfunction}(i', j') \tag{3.6}$$

In equation 3.6, g and z are the spectral and time indices of i', j', respectively, indicating the relative center offset of the frequency spectrum and time. Figure 3.3 shows the framework of the speech recognition system.

In Figure 3.3, the recognition system for speech signals mainly includes three parts: preprocessing, feature extraction and classification and recognition, in which the similarity comparison index is needed to differentiate in signal recognition. The high and low-frequency speech signals can reflect the high and low interest in the emotional content of the speech, and the long-term variation of prosodic speech can also represent the emotional difference of speech. The study introduces the prosodic feature into the statistical function to realize the transformation of the feature vector while ensuring its usability in the classifier while reducing the recognition complexity. By mixing all feature combinations and generalizing the acoustic properties of emotional speech, the most robust speech emotion feature representation set can be extracted, as shown in Figure 3.4.

Figure 3.4 shows the hybrid feature combination of speech emotion. MFCC, rhythmic and SSC features in the speech data have their own unique feature data, among which SSC features can better extract the differences in speech emotion, thus avoiding the extraction errors caused by the differences in language styles and sentence lengths.

3.2. Design of emotional speech database based on DBN - BP algorithm. Emotional features in complex environments will be affected by subjective emotional styles and relatively vague emotional demarcation

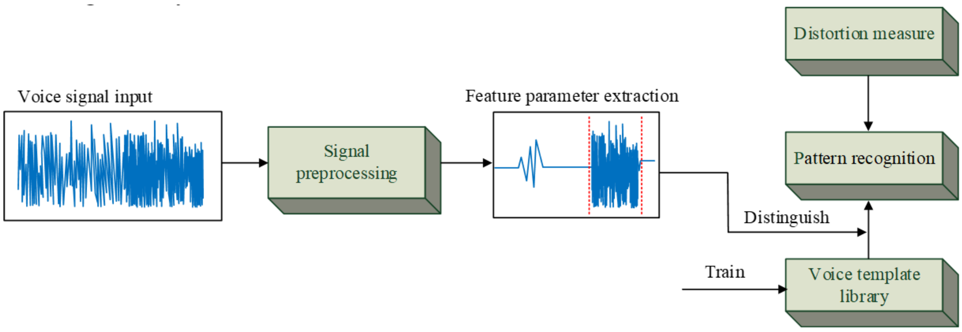


Fig. 3.3: Voice recognition system framework

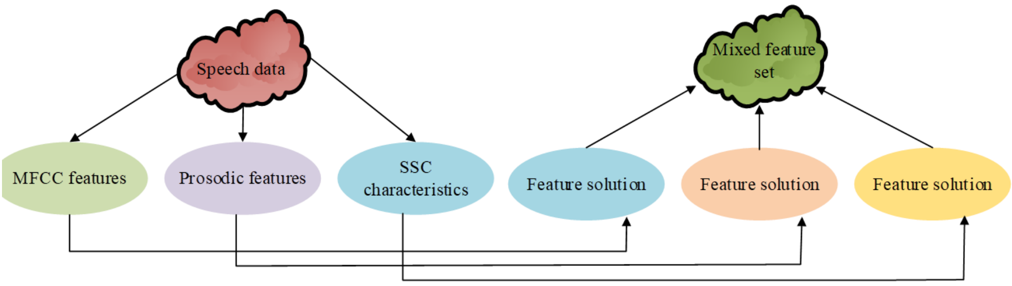


Fig. 3.4: Schematic diagram of mixed features for speech emotion extraction

points, which makes it difficult for classifiers to identify. Contextual and acoustic prosody features are highly subjective in algorithm testing. Therefore, on this basis, the deep learning algorithm is introduced to represent the speech signal at multiple levels, and the characteristic parameter factors with high robustness are extracted. Deep belief networks (DBN) are based on restricted Boltzmann machines and are trained to deal with the correlation between different hidden layers with a constructed joint distribution function. DBN includes an explicit layer responsible for data transmission and a hidden layer that adjusts the weight assignment of the data and is often trained with the " contrast divergence " algorithm, whose mathematical expression is shown in equation 3.7.

$$\begin{array}{c} P(v|h)=\prod_{i=1}^{d}P(v_i|h) \\ P(h|v)=\prod_{j=1}^{q}P(h_j|v) \end{array} \tag{3.7}$$

In equation 3.7, d, q are the neurons in the explicit and hidden layers; v and h are the state vectors corresponding to the visible and hidden layers, respectively. The updated formula of the connection weight is shown in equation 3.8.

$$\Delta w=\eta(vh^T-v'h'^T) \tag{3.8}$$

In equation 3.8, T means transposition and η denotes connection parameters. DBN is stacked and connected by multiple Boltzmann machines and effectively trained with layers. After the pre-training of each layer is completed, the whole network is trained with back propagation (BP) neural network algorithm, and a deep network model is obtained. BP algorithm realizes nonlinear transformation and learning of sample data using gradient and iterative algorithms, and its mathematical expression is shown in equation 3.9.

$$\begin{array}{c} u_j=\sum_{i=1}^{M}(\omega_{ij}x_i-\theta_j) \\ y_j=f(u_j)=\frac{1}{1+e^{-u_j}} \end{array} \tag{3.9}$$

The input value and the corresponding input signal are classified. If the corresponding input signal belongs to the input value category, it is expressed as ±1, and if not, it is 0. In equation 3.9, $x(x = 1, 2, ..., N$ is the input vector; j denotes the network node; u_j is the weighted sum of the node j threshold and the input value θ_j . The network node weights and thresholds are corrected to obtain the equation 3.10.

$$\begin{array}{c} \omega_{ij}(t+1) = \omega_{ij}(t) + \lambda\sigma_j x_i \\ \theta_j(t+1) = \theta_j(t) + \lambda\sigma_j \end{array} \tag{3.10}$$

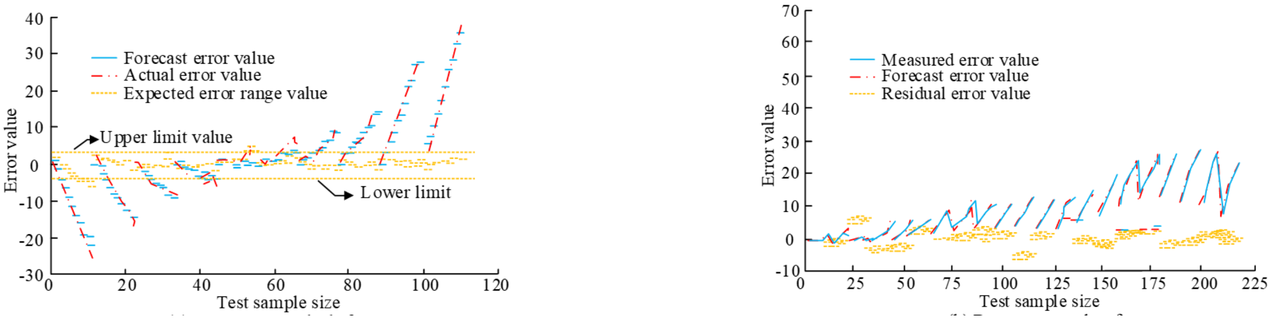
In equation 3.10, ω_{ij} is the weight x_i from node x_i to node i at a time t ; j indicates the input of the $i - th$ node; λ expresses the gain factor. Depending on whether the ideal value is clear, the value can be expressed as equation 3.11.

$$\begin{array}{c} \sigma_j = y_j(1-y_j)(d_j - y_j) \\ \sigma_j = x_i(1-x_i)\sum_k \sigma_k W_{jl} \end{array} \tag{3.11}$$

In equation 3.11, (d_j, y_j) mean the ideal output and actual output of the output node j ; l is the total number of nodes in the upper layer of the hidden layer node . When $\omega_{ij}\theta_j$ are in a steady state, the algorithm ends. Emotion recognition, as a unique feature recognition of human beings, is highly subjective, social, and cultural. Only when the two communicating parties show roughly the same emotional ups and downs, can they have the same voice characteristics. Strengthening the establishment of a voice database can effectively promote research on the characteristics of emotional data. The establishment of a voice database needs to adhere to the principles of authenticity, interactivity, continuity, and richness. Research is focused on the construction of a corpus database using specific sentence recordings and editing of related emotional video data. Taking into account the different types of emotional speech, database construction is implemented through database creation, data table definition, and data import calculation.

4. Analysis of the application results of the speech database for emotional feature extraction. The study proposed the construction of a speech database supported by hybrid algorithm based on SSC features and DBN-BP algorithm to realise the revelation of the effect of Chinese teaching on the basis of considering the characteristics of the phonological sentiment. The most important thing in Chinese teaching was to master the semantic expression between different information, in which emotion was an important acoustic feature. SSC features could be collected on the dynamic mutual information of speech signal data, while the DBN-BP algorithm extracted emotional features from speech using deep belief network and BP algorithm. This was achieved through emotional speech sample data collection, emotional feature extraction, labelling and classification of emotional categories and data storage of emotional feature information. In this study, the corpus database construction was carried out with the clips of specific utterance recordings and related emotional video data. The database construction was realized from the database creation-data table definition and data import calculation, taking into account the different types of emotional speech. The construction of emotional feature database could provide rich speech resources for international Chinese teaching, including practice materials for pronunciation, intonation and emotional expression. And through the speech database, emotional features could be used to assess and correct learners' pronunciation, helping learners to pronounce more accurately. For example, by comparing the learner's pronunciation with the standard pronunciation in the database, the learner could understand and improve his/her own pronunciation for the pronunciation characteristics of a particular emotional state. At the same time, the emotional speech samples in the database could be used to demonstrate and practice the characteristics of intonation and emotional expression in international Chinese language teaching, and learners could improve their intonation and emotional expression by imitating the emotional speech samples in the database.

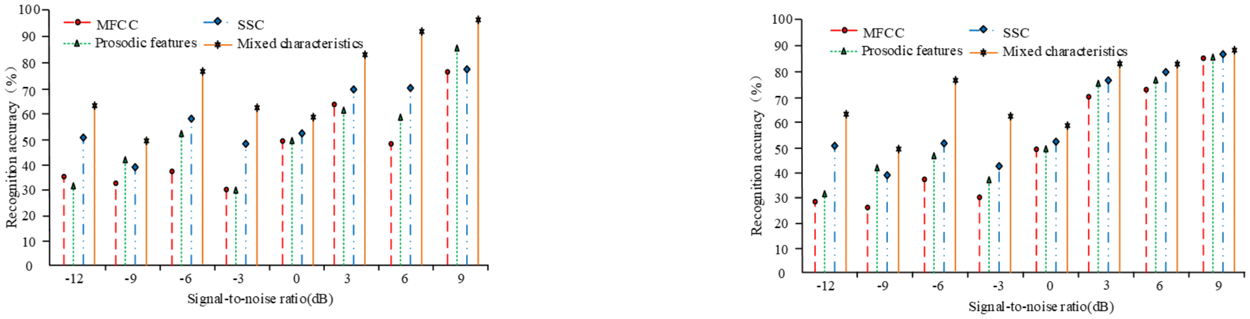
4.1. Performance test of algorithm model based on emotional features. The experimental environment was designed as follows: the central processor was Intel core i5-6500; the deep learning framework was Caffe; the interface was MATLAB; the computer memory size was 12GB; the programming language was Python. The setting of iteration times was determined based on the test and training datasets in the voice database. When testing the algorithm, it set the learning rate and the maximum number of iteration steps to 0.001 and 600, respectively, and conducted training and test analysis on the data in the speech database to



(a) Data error results before neural network compensation

(b) Data error results after neural network compensation

Fig. 4.1: Error results of data extraction before and after BP neural network compensation



(a) Differential recognition result of dynamic information

(b) Results of feature recognition accuracy under different signal-to-noise ratios

Fig. 4.2: Dynamic information recognition of different speech features and comparison of recognition accuracy results under SNR

better test the feasibility and applicability of the proposed algorithm. Figure 4.1 is an analysis of the error results of data extraction before and after adding BP neural network.

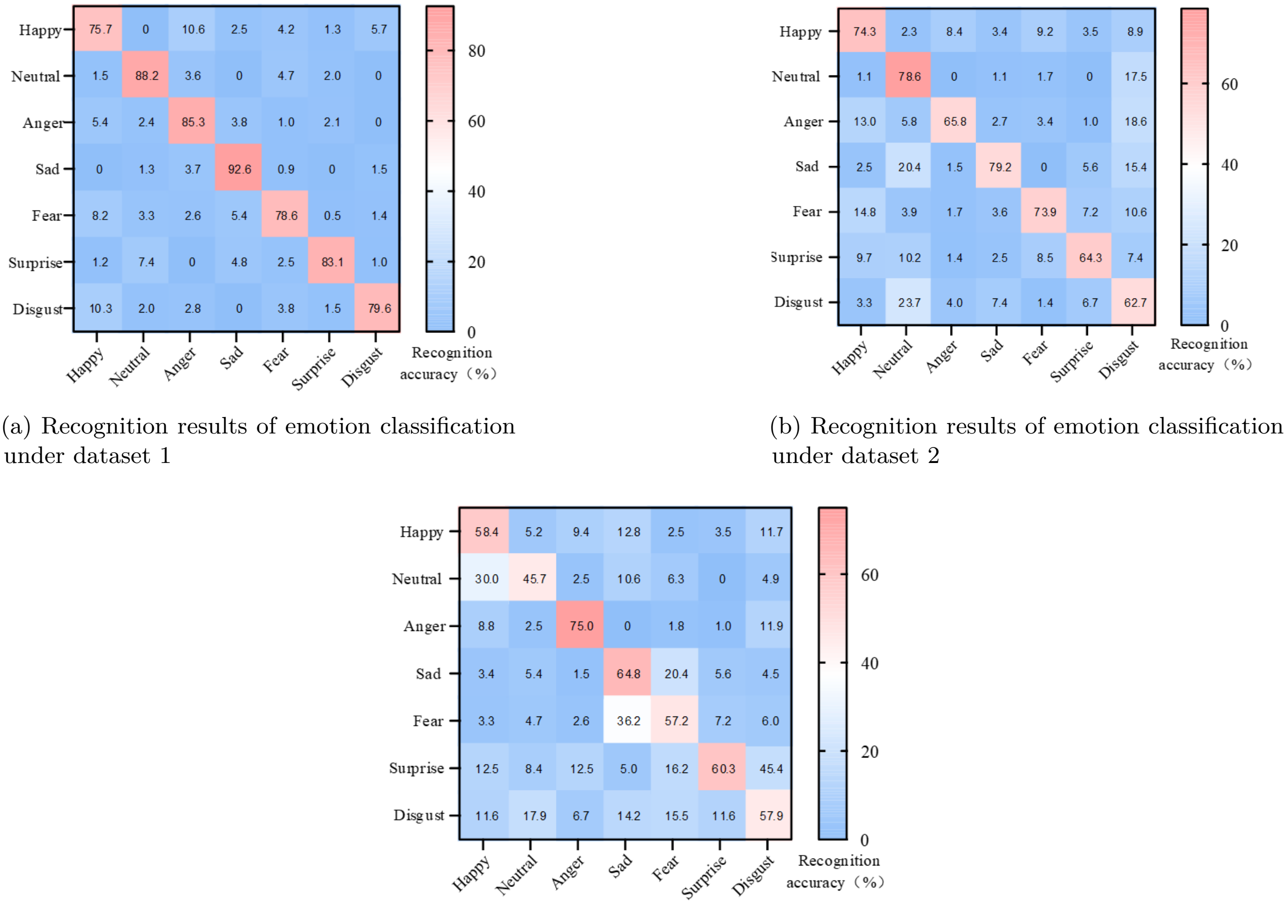
Figure 4.1 shows that the differences in the data error results were obvious before and after the data compensation with the aid of the neural network. The specific performance was that in Figure 4.1.a, the predicted and the actual error value curves shown by the proposed algorithm when extracting data features were roughly the same. The curve error fluctuation range under a small sample size was 0.38%, and the positive and negative error values were divided into two parts with the sample data volume 40 as the dividing point. The maximum prediction error value was -22 and 34, and the expected error range was between (-4, 2). The characteristic curve in Figure 4.1.b showed that the variation range between the measurement error and the prediction error was 0.042%, and the residual error value was lower than 0. The overall value was less affected by the change in the sample size, which effectively realized that the extraction of speech features ensured the accuracy of the algorithm to a certain extent. At the same time, considering that the extraction of speech signal features was more likely to be influenced by the external and objective environment and other factors, resulting in the generation of noise, the test data of the proposed feature fusion method and single speech feature extraction were compared, and the results are shown in Figure 4.2.

From Figure 4.2.a, the information accuracy rates of different speech features in differential dynamic recognition were different. When the SNR was negative, the information recognition rates from low to high were MFCC> Prosodic features>SSC>Mixed characteristics, and the smaller the SNR value, the more obvious the difference in feature information. Among them, the speech information extraction effect of fusion features was significantly higher than that of cepstral coefficient, prosodic, and SSC features. The accuracy differences at 6dB were 41.2%, 32.6%, and 19.%, respectively. And its average recognition accuracy was higher than the other three feature algorithms, and with the increase of the SNR value, the difference amplitude value decreased. The average recognition accuracy of the hybrid feature algorithm was above 94%, which was higher than that of MFCC (70.8%), Prosodic (81.3%), and SSC features (88.9%), and the maximum improvement rate exceeded 20%. The above results showed that the study of speech signal recognition from the perspective of fusion features could effectively improve its anti-noise interference ability, and had better accuracy and stability. A certain number of datasets were selected from emotional corpora in three different languages. The emotional tags expressed in the datasets were extracted, and 7 types of tags with different emotional attributes were obtained, namely happy, neutral, angry, sad, fear, surprise, and disgust were compared for recognition rates. In Figure 4.3.a - Figure 4.3.c, the languages of the three datasets were German, Chinese, and English respectively. The learning rate of the DBN network was set to 0.08 and the number of hidden layer nodes was 8000 to obtain the emotional feature recognition under different datasets, as shown in Figure 4.3.

Figure 4.3 is the accuracy confusion matrix of different sentiment label classifications, in which the row and column represent the actual and the predicted values, respectively. In the case of Figure 4.3.a dataset 1, the algorithm's recognition accuracy for 7 different sentiment attributes was 75.7%, 88.2%, 85.3%, 92.6%, 78.6%, 83.1%, 79.6%, respectively. The difference between the recognition rates of angry and happy features exceeded 15%, and the system's recognition effect of emotional features on negative attributes was always high in positive emotions. In Figure 4.3.b, under the Chinese data set, the emotion recognition rates of the algorithms were all above 60%, among which the extreme values of anger and surprise were 79.2% and 64.3%, respectively. Compared with dataset 1, the rate was worse, but its overall control over the recognition of emotional features was better. In Figure 4.3.c, the algorithm showed a recognition rate of 75% on the anger emotion feature, and the difference in recognition accuracy between the fear emotion attribute and the disgust attribute was 17.8% and 17.1%, respectively. The recognition rate of the algorithm in different datasets was not the same. The reason was that the emotional characteristics presented by different databases were related to a certain cultural background, so the recognition effects were not the same. It had good performance in emotion recognition, especially in the recognition and classification of negative emotion features.

4.2. Applicability test of algorithms based on emotional features. Paying attention to the accuracy of language information transmission and the sufficient performance of emotional expression in international Chinese teaching has effectively helped teachers improve the quality of teaching management, and to a certain extent has greatly improved teaching effectiveness. The speech data for emotion recognition proposed by the research was used to study the effect of application recognition and was combined with the BP algorithm, support vector machine algorithm (SVM), long short-term memory network (LSTM), bidirectional long short-term memory network (BILSTM). The joint attention mechanism, bidirectional long short-term memory-attention (BILSTM-Attention), was compared, and the results are shown in Table 4.1.

As shown in Table 4.1, the accuracy and recall data changes of different model algorithms under different datasets were different. The specific performance was as follows: the accuracy of the BP model under the three data sets was lower than 75%. The accuracy of the BP model decreased with the increase of the difficulty of the information covered by the data, and its performance was inferior to other comparison algorithms. The accuracy rates of the single model SVM and LSTM on the simple and medium difficulty datasets were 76.37%, 81.15% and 71.24%, 78.97%, respectively, and they also showed a certain drop in accuracy. The reason was that some informationwas missing, which in turn led to a decrease in the recognition accuracy of emotional information. The accuracy rates of the BILSTM model with the attention mechanism were 83.26% and 82.15% under the medium difficulty dataset, and the corresponding F1 values were 83.28 and 82.38. The overall performance of the data recognition was better, and its accuracy rate was only decreased by only 0.57%, significantly less than the other two single models. However, there was still a certain gap in the recognition accuracy of the DBN-BP model combined with the emotional feature analysis proposed in the study. The accuracy and recall rate of



(a) Recognition results of emotion classification under dataset 1

(b) Recognition results of emotion classification under dataset 2

(c) Recognition results of emotion classification under dataset 3

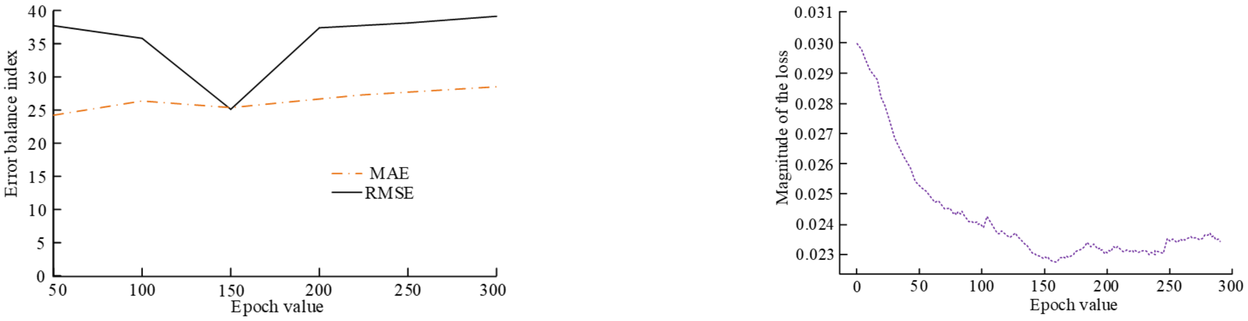
Fig. 4.3: Accuracy matrix results of emotional features on different datasets

the DBN-BP hybrid model was less affected by the difficulty of the sample data and remained at 85%. % and 86%. Its F1 values under the three data sets were 88.02, 87.65, and 86.53, respectively, and the performance of the algorithm was relatively stable. The above results indicated that the fusion DBN-BP model could effectively identify and process the features of emotional data, and its performance and application accuracy were good. The reason was that the model performed multi-dimensional processing and recognition based on the characteristics of emotional information, which overcame the problem of missing information data by a single algorithm. Then, the error analysis was carried out with the more difficult data set. Each training time was equivalent to one batch (Epoch). Multiple groups of Epochs were set to perform data statistics on the loss function results of the fusion algorithm. The results are shown in Figure 4.4.

In Figure 4.4, the evaluation index of the fusion model showed a trend of first decreasing and then increasing, and its root mean square error (RMSE) and mean absolute error (MAE) curve values tended to converge and stabilize with the increase of training batches. At the same time, the loss function value of the model algorithm showed a downward trend, and its value was lower than 0.024 in the later stage of training. The loss of data was small, and the feature retention of information data was better. The application results of the proposed model integrating emotional features and its running and test time in multiple experiments were analyzed. The average value multiple times was taken as the final result, and its emotional information extraction was

Table 4.1: Statistical results of data processing performance under different algorithms

Dataset classification	Model	Accuracy (%)	Recall rate (%)	F1
Low-difficulty data set	BP	73.25	77.13	74.38
	SVM	76.37	79.42	78.11
	LSTM	81.15	80.39	80.61
	BILSTM-Attention	83.26	84.22	83.28
	DBN-BP under emotional characteristics	85.33	87.15	88.02
Medium difficulty dataset	BP	68.22	69.13	68.34
	SVM	71.24	74.26	73.21
	LSTM	78.97	78.22	79.01
	BILSTM-Attention	82.15	81.33	82.38
	DBN-BP under emotional characteristics	86.02	88.14	87.65
Difficult data set	BP	65.23	64.31	64.08
	SVM	71.35	73.29	72.24
	LSTM	75.32	78.16	79.37
	BILSTM-Attention	79.20	81.24	80.09
	DBN-BP under emotional characteristics	85.16	86.32	86.53



(a) Training batch experiment results

(b) Experimental results of loss function

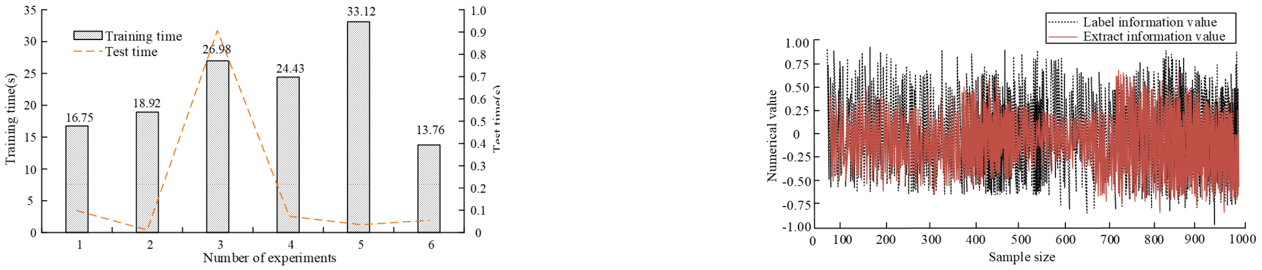
Fig. 4.4: Statistics of training batches and loss functions of mixed DBN-BP model

analyzed. The results are shown in Figure 4.5.

From Figure 4.5.a, the training time of the hybrid model under a different number of experiments was less than 1 minute, the average training time was 22.32 s. The test time of the model was less than 1s, and the maximum value was 0.93 s. The average test time was 0.64 s, and the overall test time was relatively low. It was stable when the number of experiments was greater than 4. The results in Figure 4.5.b showed that the number of samples would not cause a great interference with the performance of the model to extract information, and the information data contained in the information extraction value was basically in the label value data, and to a certain extent, the extreme value was reduced. The accuracy of identification information exceeded 86%, and the performance was good. At the same time, the validity of the algorithm was tested with the recording data of an international Chinese classroom in a university, and compared with the actual classroom emotional performance. The results are shown in Figure 4.6.

As shown in Figure 4.6, the difference between the predicted value of the model application and the real value in the four dimensions of positive, negative, neutral, and extreme emotions was small, which were 2.24%, 0.27%, 0.56%, and 0.24% respectively. The information prediction effect of emotional traits was better.

5. Conclusion. Strengthening speech emotion recognition is an important means to accelerate the promotion of intelligent human-computer interaction, which focuses on the emotional characteristics of speech data



(a) Training time and testing time of fusion model under different sample sizes

(b) Extraction of emotional feature information by fusion model

Fig. 4.5: Application time consumption of fusion model and feature extraction of emotional information

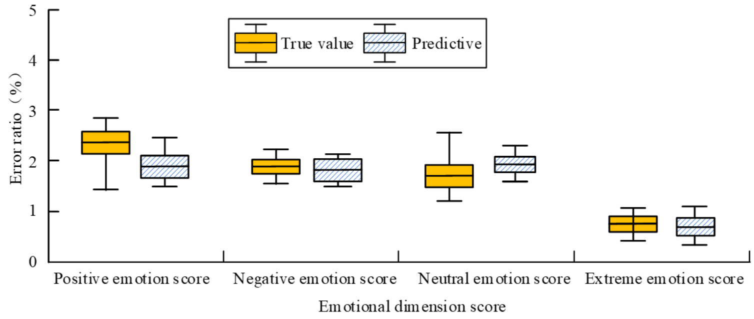


Fig. 4.6: Error comparison results between the predicted value and real value of the fusion model

in the current international Chinese teaching and effectively improves its teaching quality. The main idea of the research was to use the DBN-BP algorithm to extract emotional features, and build a speech database. After testing, the proposed algorithm showed a difference between measurement error and prediction error when extracting data features. The range of change was 0.042%, and the accuracy of dynamic identification information was higher than that of cepstral coefficient (70.8%), prosodic (81.3%) and SSC features (88.9%). At the same time, the recognition rate of different emotions of the DBN-BP model in the Chinese database was above 60%, the overall recognition control was good, and the accuracy and recall rate shown in the comparison with other algorithms were less affected by samples. The influence of the difficulty of the data remained above 85% and 86%, and the F1 values of the corresponding data sets were 88.02, 87.65, and 86.53, respectively, which were higher than the performance of other algorithms of the same dimension. The value of the model proposed was lower than 0.024 in the later stage of training, and the feature retention of information data was better. The average training time and test time were 22.32s and 0.64s, and the accuracy rate of emotional label recognition information exceeded 86%. The difference in scores on all four emotional dimensions was less than 3 percent. Focusing on emotional feature extraction can effectively improve the accuracy and applicability of speech recognition, and strengthening the multi-dimensional inspection of data is one of the ideas for future research and improvement.

Fundings. The research is supported by the Key Research Base of Humanities and Social Sciences of Sichuan Province-Sichuan International Education Development Research Center“Research on University Alliance Serving the Construction of the ‘the Belt and Road’(SCGJ2022-20); Foundation project: Panzhihua University 2022 school-level research project: Ideas and opportunities for the development of Chinese international education in sino-foreign cultural exchanges.

REFERENCES

[1] Mulvey, B. International higher education and public diplomacy: A case study of Ugandan graduates from Chinese universities. *Higher Education Policy.* **33**, 459-477 (2020)

[2] Kaur, J., Singh, A. & Kadyan, V. Automatic speech recognition system for tonal languages: State-of-the-art survey. *Archives Of Computational Methods In Engineering.* **28**, 1039-1068 (2021)

[3] Zehra, W., Javed, A., Jalil, Z. & Others Cross corpus multi-lingual speech emotion recognition using ensemble learning. *Complex & Intelligent Systems.* **7**, 1845-1854 (2021)

[4] Fan, X., Zhao, H., Chen, X. & Others Deceptive Chinese speech detection based on sparse decomposition of cepstral feature. *Chinese Journal Of Acoustics.* **38** pp. 01 (2019)

[5] Pan, L., Hu, L. & Li, Z. Simulation of English part-of-speech recognition based on machine learning prediction algorithm. *Journal Of Intelligent & Fuzzy Systems: Applications In Engineering And Technology.* **40**, 2409-2419 (2021)

[6] Mnassri, A., Cherif, A. & Bennasr, M. Algorithm Optimizing SVM Multi-Class Kernel Parameters Applied in Arabic Speech Recognition. *International Journal Of Systems Signal Control & Engineering Applications.* **12**, 85-92 (2019)

[7] Wang, S., Zhou, P., Chen, W. & Others Exploring run-transducer for Chinese speech recognition//2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC). *IEEE.* pp. 1364-1369 (2019)

[8] Koduru, A., Valiveti, H. & Budati, A. Feature extraction algorithms to improve the speech emotion recognition rate. *International Journal Of Speech Technology.* **23**, 45-55 (2020)

[9] Kumaran, U., Radha Rammohan, S., Nagarajan, S. & Others Fusion of Mel and gammatone frequency cepstral coefficients for speech emotion recognition using deep C-RNN. *International Journal Of Speech Technology.* **24**, 303-314 (2021)

[10] Kerkeni, L., Serrestou, Y., Raoof, K. & Others Automatic speech emotion recognition using an optimal combination of features based on EMD-TKEO. *Speech Communication.* **114** pp. 22-35 (2019)

[11] Huang, Y., Tian, K., Wu, A. & Others Feature fusion methods research based on deep belief networks for speech emotion recognition under noise condition. *Journal Of Ambient Intelligence And Humanized Computing.* **10** pp. 1787-1798 (2019)

[12] Daneshfar, F. & Kabudian, S. Speech emotion recognition using discriminative dimension reduction by employing a modified quantum-behaved particle swarm optimization algorithm. *Multimedia Tools And Applications.* **79**, 1261-1289 (2020)

[13] Widodo, H., Fang, F. & Elyas, T. The construction of language teacher professional identity in the Global Englishes territory:'we are legitimate language teachers'. *Asian Englishes.* **22**, 309-316 (2020)

[14] Chen, M., He, X., Yang, J. & Others 3-D convolutional recurrent neural networks with attention model for speech emotion recognition. *IEEE Signal Processing Letters.* **25**, 1440-1444 (2018)

[15] Kwon, S. Optimal feature selection based speech emotion recognition using two-stream deep convolutional neural network. *International Journal Of Intelligent Systems.* **36**, 5116-5135 (2021)

[16] Widodo, H., Fang, F. & Elyas, T. The construction of language teacher professional identity in the Global Englishes territory: 'we are legitimate language teachers'. *Asian Englishes.* **22**, 309-316 (2020)

[17] Yuan, R., Li, S. & Yu, B. Neither "local" nor "global": Chinese university students' identity paradoxes in the internationalization of higher education. *Higher Education.* **77**, 963-978 (2019)

[18] Xinhan, N. Intelligent analysis of classroom student state based on neural network algorithm and emotional feature recognition. *Journal Of Intelligent And Fuzzy Systems.* **40**, 1-12 (2020)

[19] Hu, J. & Zhang, H. Recognition of classroom student state features based on deep learning algorithms and machine learning. *Journal Of Intelligent & Fuzzy Systems: Applications In Engineering And Technology.* **40** pp. 2 (2021)

[20] Byun, S. Lee S P . *Study On A Speech Emotion Recognition System With Effective Acoustic Features Using Deep Learning Algorithms.* **2021** pp. 4 (0)

[21] Shah, V. Mehta M .Emotional state recognition from text data using machine learning and deep learning algorithm. *Concurrency And Computation: Practice And Experience.* **2022** pp. 17 (0)

[22] Atmaja, B., Sasou, A. & Akagi, M. Survey on bimodal speech emotion recognition from acoustic and linguistic information fusion. *Speech Communication.* **140** pp. 11-28 (2022)

[23] Pham, N. Dang N M D, Nguyen S D. A Method upon Deep Learning for Speech Emotion Recognition. *Journal Of Advanced Engineering And Computation.* **4** pp. 4 (2021)

[24] Long, L. & Liang, T. Multi-Distributed Speech Emotion Recognition Based on Mel Frequency Cepstogram and Parameter Transfer. (Chinese Journal,0)

Edited by: Mudasir Mohd
Special issue on: Scalable Computing in Online and Blended Learning Environments: Challenges and Solutions
Received: May 16, 2023
Accepted: Nov 16, 2023