An Analysis of Emotional Responses of Students in Bilingual Classes and Adjustment Strategies

https://doi.org/10.3991/ijet.v18i01.37125

Suyun Wen(✉)
Department of Public Foreign Languages, Shijiazhuang University of Applied Technology, Shijiazhuang, China
2003110496@sjzpt.edu.cn

Abstract—Students' willingness to participate in bilingual communication is greatly influenced by their positive emotions in the bilingual class. The automatic recognition of students' emotional state in bilingual class can assist teachers to correctly master the laws of students' emotional changes during the bilingual learning process as fast as possible. However, the speech emotion features extracted by existing speech emotion recognition models are not universal and not suitable for bilingual speech emotion recognition, and the accuracy needs to be improved. To cope with these issues, this paper aims to study the emotional responses of students in bilingual class based on emotional analysis and provide adjustment strategies for it. At first, the signals of students' speech records in bilingual class were pre-processed, the one-dimensional features of the signals were converted into 2-dimensional speech spectrum features so as to attain more useful information to facilitate the emotional recognition of students' speech records in the bilingual class. Then this paper combined the bi-linear Convolution Neural Network (CNN) with capsule network to construct a bilingual class student emotion recognition model, and experimental results verified the effectiveness of the constructed model.

Keywords—emotional analysis, bilingual class, emotion recognition, bi-linear convolution neural network (CNN), capsule network

1 Introduction

As positive psychology has been applied in the field of education, now teachers have paid more attention to students' positive emotions in the class, especially the bilingual teachers who need to interact with students more often [1–6]. Studies have shown that students' willingness to engage in bilingual communication is significantly influenced by their positive emotions in the bilingual class [7–16]. For this reason, world field scholars have conducted research on students' emotional responses in bilingual class, and found that the automatic recognition of students' emotional state in bilingual class can assist teachers to correctly master the laws of students' emotional changes during the bilingual learning process as fast as possible, thereby formulating targeted teaching intervention measures according to the actual emotional state of students [17–19].

Liang et al. [20] pointed out that speech emotion recognition is one of the research hotspots in artificial intelligence. In the paper, the authors attempted to introduce speech emotion into classroom teaching and extract the emotional features of speech, they proposed a multi-channel convolution combining with SEnet network as an emotion recognition model, which performed well in terms accuracy, F1 value, and recall rate on a self-built emotion dataset. Li et al. [21] proposed a DNN-based multi-modal learning emotion analysis method, which integrated video and voice to detect the real-time learning emotions of students, the authors applied this method to the automatic recognition of the learning emotions of students in primary school English class, and used a PAD emotion scale to correspond learning emotions to learning states, then teachers could judge students' learning state according to the changes in their learning emotions and adjust the teaching methods and strategies in time. He et al. [22] used deep learning and attention mechanism methods to automatically modelling the temporal process, and built a DMTN-BTA model based on multitask learning for recognizing students' emotions through classroom videos. The proposed method contained a CNN used for spatio-temporal feature extraction, and a BLSTM-RNN used for emotion recognition which had introduced a novel bridge-temporal-attention. Lei et al. [23] proposed a deep convolution classroom facial analysis method based on channel interaction for the purpose of solving problems in camera shooting distance, indoor illumination, and other classroom scene factors, and the unbalanced expression image and poor recognition effect caused by fuzzy facial spectra of students. Their experimental results showed that data enhancement and model improvement have good effects on facial expression analysis. Liang et al. [24] studied the teacher speech signals and designed a set of emotion detection audio processing system to judge their emotions based on the speech. The authors used the recurrent neural network algorithm to build a speech emotion recognition classification model, pre-processed the data based on pre-weighting, frame-adding window, and endpoint detection, re-classified the emotions, and established a speech emotion corpus of teacher evaluation system. Their experimental results showed that, the improved eigenvalues of Mel frequency cepstral coefficients and neural networks can improve the recognition rate of speech emotions more effectively than conventional speech emotion recognition methods, and can be applied to speech emotion recognition in classroom teaching.

For bilingual speech emotion recognition, the existing corpus is not individualized enough, the speech emotion features extracted by existing speech emotion recognition models are not universal and not suitable for bilingual speech emotion recognition, and the accuracy needs to be improved. Speech emotion recognition models with excellent performance are generally improved or build based on fine-tuned conventional models, so choosing the right combination model is an important condition for effective emotion recognition of bilingual speech. This paper studied the emotional responses of students in bilingual class and the adjustment strategies. In the second chapter, this paper preprocessed the signals of students' voice records in bilingual class to make the signals smoother so that the features of the signals could be extracted easier. In the third chapter, this paper converted the one-dimensional features of recording signals into 2-dimensional speech spectrum features so as to attain more useful information and to facilitate the recognition of students' speech emotions in bilingual class. In the fourth

chapter, this paper combined the bi-linear CNN with capsule network to construct a combination model for recognizing emotions of students in bilingual class. At last, experimental results verified the effectiveness of the constructed model.

2 Preprocessing of recording signals of bilingual class

During the recording process, there might be various interference factors such as environmental noise, equipment noise, speaking interval, low clarity, and irregular duration in the record database of bilingual class. To make the signals of students' voice records in bilingual class smoother to facilitate feature extraction, the original signals of bilingual speech records need to be pre-processed.

In the bilingual speech records of students, the energy of high-frequency areas was less than the energy of the low-frequency areas. This paper used the digital high-pass filter to pre-accentuate the recording signals to increase the signal-to-noise ratio of the bilingual speech records and enhance the energy of the high-frequency areas. Assuming: $Y[L]$ represents the input signal sequence of the bilingual speech records; $X[L]$ represents the output signal sequence, then the formula below gives the formula of filtering processing:

$$X[L] = Y[L]*0.97Y[L-1] \tag{1}$$

The long-time recording signals need to be truncated, that is, the overlapping area between a frame and the next frame needs to be separated based on its macro stability, and here the length of the overlapping area is defined as the frame shift. In this paper, the frame length and frame shift were respectively set as 25 and 10 ms so as to maximize the retained information of student emotions in the bilingual speech records.

The preprocessed recording signals must be continuous, so they need to be superimposed with the window function, namely to be subjected to the windowing operation. The window function can be considered as a data of equal length, and its expression is given by the following formula:

$$Q(m) = \begin{cases} 0.54 - 0.46\cos\left(\dfrac{2\pi m}{M-1}\right), 0 \le m \le M-1 \\ 0, Others \end{cases} \tag{2}$$

To get the frequency spectrum of the signals of the bilingual speech records, the fast Fourier transform was applied to the recording signals after windowing. Assuming: $B(k)$ represents the frequency domain sample of the recording signals, $b(n)$ represents the time domain sample, N represents the size of the fast Fourier transform, expression of the fast Fourier transform is given by the following formula:

$$B(k) = \sum_{n=0}^{N-1} b(n) p^{-i\left(\frac{2\pi}{N}\right)nk} (k = 0, 1, L\ N-1) \tag{3}$$

3 Feature extraction of speech spectrum of bilingual record signals

In this paper, the one-dimensional features of the recoding signals were converted into two-dimensional speech spectrum features to attain more useful information for the recognition of student emotions in the bilingual class. Compared with conventional CNN, the bi-linear CNN with two parallel convolution layers can perform deep extraction on the texture features of the spectra corresponding to the recording signals through the outer product operation of different recording signal features, thereby realizing more satisfactory emotion recognition accuracy. There're certain similarities in the features of different recoding signals, to solve the unsatisfactory emotion recognition accuracy caused by such similarities, this paper introduced the capsule network based on bi-linear CNN.

Generally, the speech spectra are of large size, so this paper converted the large-size speech spectrum into the Mel spectrogram based on the Meyer filter group to ensure that the bi-linear CNN could attain speech spectrum input with reasonable size.

Features of the Mel spectrogram could be attained by the following steps:

Step 1: Perform a series of preprocessing operations such as filtering, framing, windowing, and Fourier transform on the original signals of bilingual speech records of students;

Step 2: Calculate the Mel frequency;

Step 3: Perform multiplying and adding operations on the corresponding energy spectrum to obtain the spectrogram. Some research results have verified that the Mel frequency is more consistent with the auditory characteristics of human, namely the two have a basically linear relationship, its expression is given by Formula 4:

$$g_{MF} = 2595 \cdot \lg\left(1 + \frac{g}{700Hz}\right) \tag{4}$$

Formula 5 gives the expression for converting into actual frequency:

$$NS^{-1}(n) = 700\left(p^{\frac{n}{1125}} - 1\right) \tag{5}$$

Assuming: L represents the length of FFT, Gr represents the sampling rate, $\Gamma()$ represents the integer function, the formula below calculates the Mel frequency resolution:

$$g(i) = \Gamma\left(\frac{(L+1) * f(i)}{Gr}\right) \tag{6}$$

Assuming: n represents the n-th Mel filter; $g(n–1)$ and $g(n+1)$ represent the upper frequency and lower frequency of the filter; $g(n)$ represents the center frequency of the Mel filter; l represents the serial number of dots; multiple filters are defined and denoted as $F_n(l)$, the formula for calculating the output of the filter is:

$$F_n(l)=\begin{cases}\dfrac{l-g(n-1)}{g(n)-g(n-1)}, g(n-1)\le l\le g(n)\\ \dfrac{l-g(n-1)}{g(n)-g(n-1)}, g(n)\le l\le g(n+1)\\ \\ 0, \textit{Others}\end{cases} \tag{7}$$

Assuming: $|a(l)|^2$ represents the size of the energy of the l-th dot in the energy spectrum, the Mel spectrogram was attained by multiplying and adding the energy spectrum with the filter output corresponding to each dot, the calculation formula of the Mel spectrogram is given by the following formula:

$$MclSpec(n)=\sum_{l=g(n-1)}^{g(n+1)} F_n(l)*\left|A(l)\right|^2 \tag{8}$$

4 Construction of the bilingual class student emotion recognition model

In this paper, bi-linear CNN was adopted to extract the fine-grained features of the spectra corresponding to the recording signals of student speech in the bilingual class, the effect of spectrum texture recognition was good, which can be used for the emotion classification and recognition of different language speech signals. Figure 1 gives the structure of bi-linear CNN. Assuming: g_x and g_y represent the feature functions extracted by different convolution kernels; E represents the pooling operation; D represents the speech signal emotion classification function, then the bi-linear CNN model can be written as G=(g_x, g_y, E, D).

Feature function $g(h)$ can map spectrum F corresponding to the input recording signals and its position K to into a feature with a size of $d\times C$, which satisfies the function mapping relationship of g:$K\times F\rightarrow R^{d\times C}$. In order to attain the effective bi-linear convolution features, the different features of spectra corresponding to the input recording signals at a same position were combined through the outer product operation of the matrix. Assuming: Y represents the bi-linear operation, then there is:

$$Y(k, I, g_x, g_y)=g_x(k, I)^T g_y(k, I) \tag{9}$$

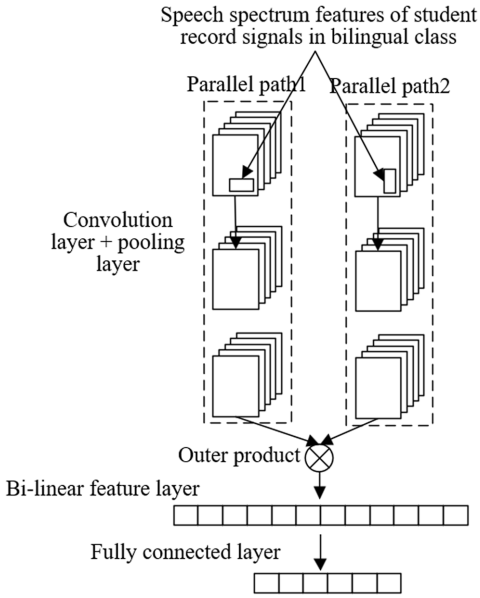


Fig. 1. Structure of bi-linear CNN

Because for different types of languages, there will be large differences in the expression of a same emotion, so when recognizing the emotions in the speech signals containing two kinds of languages, there might be differences in the extracted features in terms of size, direction, and state. Although the convolution operation of CNN can extract more subtle matrix features based on the attained emotion features of recording signals, it cannot deal with the changes in above features, and the emotion recognition effect will be greatly affected, so in order to make up for the defect of CNN, this paper introduced the capsule network.

Capsules in the capsule network transmit in the form of vectors, the probability and attribute of the emotion features of recording signals are respectively characterized by the length and dimension values of the vectors. For the recording signals being detected, if the corresponding spectrum and its position or state undergoes changes, then it can be considered that only the vector direction has changed, and the vector length hasn't changed.

Assuming: q_i represents the transition matrix, then the vector of the extracted fine-grained features of the spectrum corresponding to the recoding signals was multiplied by q_i to get the vector that can be recognized by the capsules in the capsule network, the formula is:

$$v_i = q_i x_i \tag{10}$$

Different capsule vectors need to be assigned with different weights, the expression of weight d_i is given by the following formula:

$$d_i = softmax\,(y_i) \tag{11}$$

The weighted vector r can be expressed as:

$$r = \sum_i d_i v_i \tag{12}$$

Because the value range of the probability of emotion features of recording signals is [0,1], this paper introduced a squashing function to compress the capsule vector length that represents this parameter, and the formula is:

$$u_i = \frac{\|r\|^2}{1+\|r\|^2}\frac{r}{\|r\|} \tag{13}$$

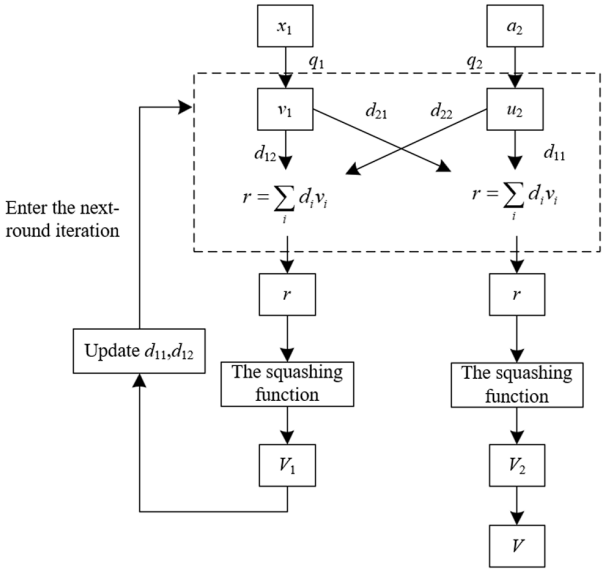


Fig. 2. Flow of the dynamic routing algorithm

The calculation results of the above formula were updated based on the dynamic routing algorithm. Figure 2 shows the flow of the dynamic routing algorithm. If the algorithm terminates, then the final result u_i is output; if the algorithm does not terminate, then y_i is updated, and the algorithm enters the next-round iteration, the following formula gives the calculation formula of y_i:

$$y_i^* = y_i + v_i u_i \tag{14}$$

When a certain emotion feature appears in the feature spectrum corresponding to the recording signals, then the final output of the capsule network is an instantiated vector that represents the probability of the occurrence of the emotion feature, and the length of the vector should be relatively long. The following formula gives the expression of edge loss function K_l used for student emotion recognition:

$$K_l = O_l max\left(0, n^+ - \|u\|^2\right) + \mu(1 - O_l) max\left(0, \|u\| - n^-\right)^2 \tag{15}$$

If a student emotion is recognized as the l-th type, then O_l=1; otherwise, O_l=0. n^+ and n^- were used to limit the length of the vector, and their values were 0.9 and 0.1. In order to reduce the loss of recognition when the student emotion type does not exist, μ was set to 0.5.

After the speech spectrum features corresponding to the recoding signals of students in bilingual class were constructed, this paper combined the bi-linear CNN with capsule network to construct a combination model, the bilingual class student emotion recognition model, as shown in Figure 3, the bi-linear CNN can further attain the features of the speech spectrum textures corresponding to the recoding signals to improve the accuracy of emotion recognition, and the capsule network had made up for the defect that the CNN cannot deal with the changes of features in size, direction, and state.

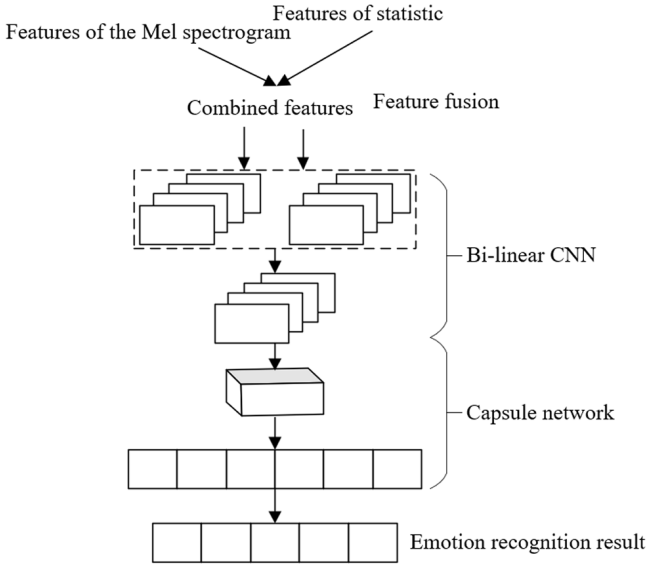


Fig. 3. Structure of the combination model of bilingual class student emotion recognition

The features of the Mel spectrogram corresponding to the recording signals of student speech in bilingual class and the features of statistic were fused to generate the combined features with time series attribute. The two parallel paths in the bi-linear CNN then performed convolution and pooling operations on the combined features for multiple times, and the deep-level features of the recording signals were attained. Next, after the outer product processing, the deep-level features were input into the capsule network, and eigenvectors representing the probability and attribute of the emotion features of the recording signals were judged to attain the ultimate correct recognition results of the emotions of students in the bilingual class.

The fusion of the features of the Mel spectrogram corresponding to the recording signals of student speech in bilingual class and the features of statistic has a large impact on the quality of the extracted features. The specific steps of feature fusion are:

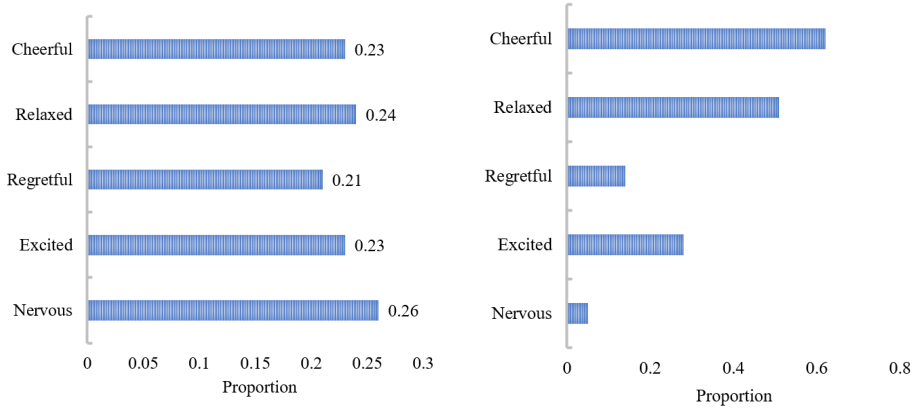
Step 1: The recording signals were converted into a Mel spectrogram feature matrix with a size of $m*n$, wherein m and n are respectively the frequency length and time length of the matrix.

Step 2: The features of the frame-statistic of the recoding signals were extracted to attain a m_1*n feature matrix, wherein m_1 and n respectively correspond to the feature number and time length.

The feature matrix of the Mel spectrogram and the feature matrix of the statistic were fused according to time series to generate a combined feature matrix with a dimension of $(m+m_1)*n$. Assuming: r_{ij} and K_{ij} respectively represent the vector of the j-th segment of the i-th Mel spectrogram and the i-th frame-statistic feature, then there is:

$$w_{ij} = [r_{ij}, K_{ij}] \tag{16}$$

5 Experimental results and analysis



(a) Training set **(b) Test set**

Fig. 4. Distribution of emotion type samples of the training set and test set

Proportions of the emotions in the training sample set and test sample set of the recording signals of students in the bilingual class are given in Figure 4a and b, the emotions contain five types: cheerful, relaxed, regretful, excited, and nervous.

Table 1. Student emotion recognition results of different networks

Emotion Recognition Model	*IPL* Features	Mel Spectrogram Features	Statistic Features	Combined Features
VGG16	75.62%	71.92%	71.6%	75.4%
Resnet34	70.9%	71.4%	77.9%	80.9%
SVM	74.6%	77.3%	70.7%	76.2%
LSTM	79.3%	76.9%	73.4%	73.4%
Residual network	85.7%	83.1%	88.7%	85.9%
The proposed model	80.4%	87.4%	89.4%	92.7%

Table 2. Confusion matrix of student emotion recognition numbers corresponding to Mel spectrogram features

	Cheerful	Relaxed	Regretful	Excited	Nervous
Cheerful	36	2	1	4	7
Relaxed	2	25	8	3	5
Regretful	4	1	17	2	4
Excited	7	2	6	28	6
Nervous	2	9	1	6	29

Table 3. Confusion matrix of student emotion recognition numbers corresponding to combined features

	Cheerful	Relaxed	Regretful	Excited	Nervous
Cheerful	36	2	7	4	5
Relaxed	6	22	9	5	1
Regretful	8	5	28	3	7
Excited	3	8	6	29	4
Nervous	8	4	2	1	26

The final student remotion recognition results are shown in Table 1. According to the table, compared with other five types of emotion recognition models, the proposed model that fused the features of the Mel spectrogram and the frame-statistic exhibited better recognition effect, and it can further improve the accuracy of emotion recognition of different students in the bilingual classroom environment.

To further verify the effectiveness of the feature fusion processing of the Mel spectrogram and the frame-statistic, this paper designed comparative experiments, and Tables 2 and 3 respectively give the confusion matrix of student emotion recognition numbers corresponding to Mel spectrogram features, and the confusion matrix of student emotion recognition numbers corresponding to combined features. By comparing the two tables, it's known that the combined features can effectively and correctly classify the five types of student emotions in the bilingual class. More intuitively, according to the comparison of student emotion recognition rate for different features shown in Figure 5, the combined features had improved the recognition rate of the five types of student emotions to a certain extent.

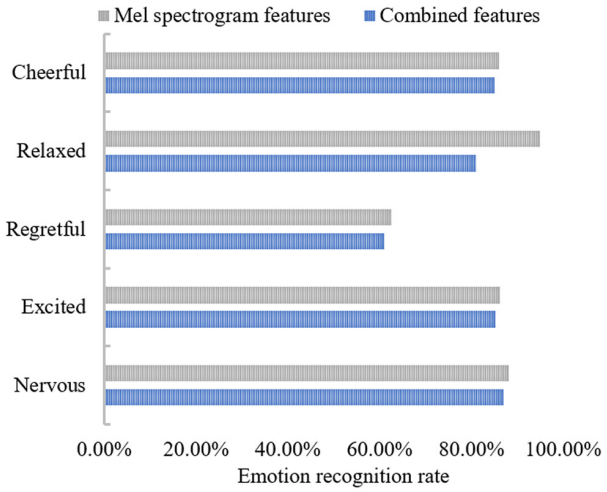
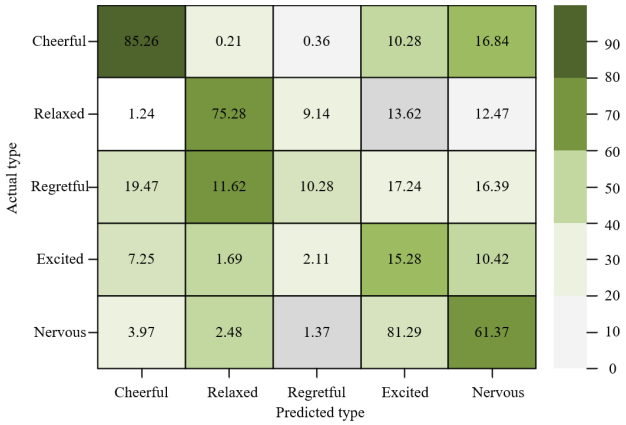
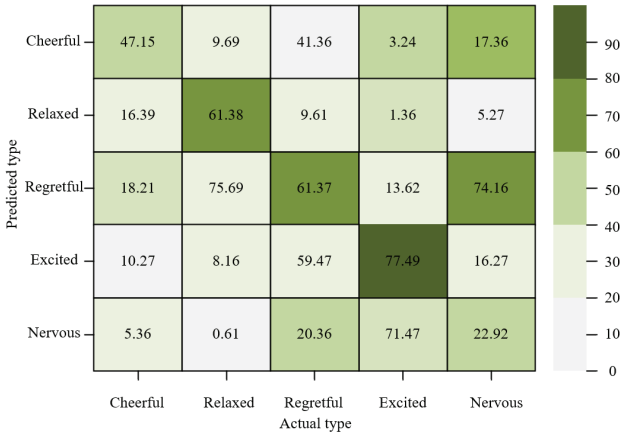


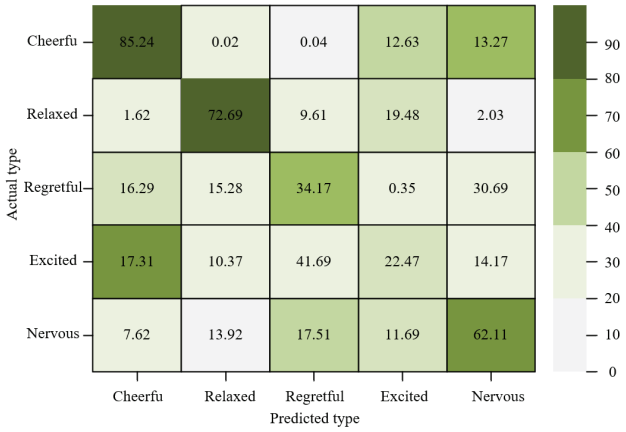
Fig. 5. Comparison of student emotion recognition rate for different features



(a) English



(b) English+Chinese



(c) German+Chinese

Fig. 6. Confusion matrixes of speech emotion recognition model under monolingual and bilingual conditions

To verify the applicability of the bilingual class student emotion recognition model constructed in this paper to different language samples, this paper took the English, English+Chinese, and German+Chinese classes as examples to carry out related experiments, and Figure 6 gives the experimental results. By observing the three confusion matrixes corresponding to English, English+Chinese, and German+Chinese classes, we can know that, based on the extracted features of bilingual speech signal spectra, the proposed model achieved ideal recognition rate of student emotions, that is, the model can well capture the features of emotion factors that are not significant enough, which had further verified the effectiveness of the proposed model. Based on the output of this model, teachers can adjust their teaching strategies in time to promote the interaction between teachers and students in bilingual class.

6 Conclusion

This paper studied emotional responses of students in bilingual class based on emotional analysis and the adjustment strategies. At first, this paper preprocessed the recording signals of bilingual class and converted the one-dimensional features of the recording signals into 2-dimensional speech spectrum features to attain more useful information and to facilitate the recognition of student emotions in the bilingual class. Then, this paper combined the bi-linear CNN with capsule network to construct a combination model for recognizing emotions of students in the bilingual class. After that, the experimental results of student emotion recognition of different networks proved that the proposed model outperformed other five emotion recognition models in terms of recognition effect. The confusion matrix of student emotion recognition numbers corresponding to Mel spectrogram features, and the confusion matrix of student emotion recognition numbers corresponding to combined features were given, which had further verified the effectiveness of the feature fusion processing of the Mel spectrogram and the frame-statistic. At last, the confusion matrixes of speech emotion recognition model under monolingual and bilingual conditions were given, which had verified the applicability of the bilingual class student emotion recognition model constructed in this paper to different language samples.

7 References

[1] Walifa, R.K. (2020). The effect of stressful factors, locus of control and age on emotional labour and burnout among further and adult education teachers in the U.K. International Journal of Emerging Technologies in Learning, 15(24): 26–37. https://doi.org/10.3991/ijet.v15i24.19305

[2] Walifa, R.K. (2020). The influence of policy on emotional labour and burnout among further and adult education teachers in the U.K. International Journal of Emerging Technologies in Learning, 15(24): 232–241. https://doi.org/10.3991/ijet.v15i24.19307

[3] Wang, S.Y. (2021). Online learning behavior analysis based on image emotion recognition. Traitement du Signal, 38(3): 865–873. https://doi.org/10.1134/S105466182103024X

[4] Cui, L.L., Kong, W.G., Sun, Y.M., Shao, L. (2022). Expression identification and emotional classification of students in job interviews based on image processing. Traitement du Signal, 39(2): 651–658. https://doi.org/10.18280/ts.390228

[5] Zhang, Z., Li, Y., Sun, S., Tang, Z. (2022). Intervention effect of group counseling based on positive psychology on psychological crisis of college student. Computational Intelligence and Neuroscience, 2022: 3132016. https://doi.org/10.1155/2022/3132016

[6] Jiang, L., Zhang, Y. (2022). Prediction and influencing factors of big data on college students' positive psychological quality under mobile wireless network. Mobile Information Systems, 2022: 4344747. https://doi.org/10.1155/2022/4344747

[7] Yao, R. (2021). Construction of bilingual teaching hierarchy model based on flipping classroom concept. In International Conference on Data and Information in Online, Hangzhou, China, pp. 421–425. https://doi.org/10.1007/978-3-030-77417-2_34

[8] Guo, J., Cong, X., Miao, Z., Feng, Z., Yang, E. (2021). Research and practice on bilingual teaching method of applied talents training based on flipped classroom teaching mode. In 2021 2nd International Conference on Big Data and Informatization Education (ICBDIE), Hangzhou, China, pp. 591–594. https://doi.org/10.1109/ICBDIE52740.2021.00140

[9] Palacios-Hidalgo, F.J. (2020). Video-based analysis of pre-service primary bilingual teachers' perceptions about the inclusion of gender and LGBT+ issues in the EFL classroom. In 2020 Sixth International Conference on E-Learning (Econf), Sakheer, Bahrain, pp. 110–114. https://doi.org/10.1109/econf51404.2020.9385471

[10] Xia, W., Zhang, Z., Guo, C. (2019). Novel education technology may derive from personal genome data: A language gene polymorphism site potentially associated with translation-writing errors in a bilingual classroom of Chinese students. In Proceedings of the 2019 International Conference on Modern Educational Technology, New York, United States, pp. 45–48. https://doi.org/10.1145/3341042.3341067

[11] Suárez, E., Otero, V. (2014). Leveraging the cultural practices of science for making classroom discourse accessible to emerging bilingual students. Boulder, CO: International Society of the Learning Sciences, 2: 800–807. https://doi.dx.org/10.22318/icls2014.800

[12] Ren, X., Han, J., Hu, D. (2011, August). Reform and Practice on Bilingual Classroom Instruction Patterns in Professional Courses Based on Web Resources. In International Conference on Computer Science, Environment, Ecoinformatics, and Education, Wuhan, China, pp. 122–126. https://doi.org/10.1007/978-3-642-23339-5_22

[13] Zhang, L. (2010). Classroom bilingual teaching quality standards system based on Delphi method and AHP. In 2010 International Conference on E-Health Networking Digital Ecosystems and Technologies (EDT), Shenzhen, China, pp. 152–155. https://doi.org/10.1109/EDT.2010.5496488

[14] Van Laere, E., Agirdag, O., van Braak, J. (2016). Supporting science learning in linguistically diverse classrooms: Factors related to the use of bilingual content in a computer -based learning environment. Computers in Human Behavior, 57: 428–441. https://doi.org/10.1016/j.chb.2015.12.056

[15] Bovo, R., Callegari, E. (2009). Effects of classroom noise on the speech perception of bilingual children learning in their second language: Preliminary results. Audiological Medicine, 7(4): 226–232. https://doi.org/10.3109/16513860903189499

[16] Bourguet, M.L. (2006). Introducing strong forms of bilingual education in the mainstream classroom: A case for technology. In Sixth IEEE International Conference on Advanced Learning Technologies (ICALT'06), Kerkrade, Netherlands, pp. 642–646. https://doi.org/10.1109/ICALT.2006.1652523

[17] Su, C., Wang, G. (2020). Design and application of learner emotion recognition for classroom. In Journal of Physics: Conference Series, 1651(1): 012158. https://doi.org/10.1088/1742-6596/1651/1/012158

[18] Liang, Y. (2019). Intelligent emotion evaluation method of classroom teaching based on expression recognition. International Journal of Emerging Technologies in Learning, 14(4): 127–141. https://doi.org/10.3991/ijet.v14i04.10130

[19] Putra, W.B., Arifin, F. (2019). Real-time emotion recognition system to monitor student's mood in a classroom. In Journal of Physics: Conference Series, 1413(1): 012021. https://doi.org/10.1088/1742-6596/1413/1/012021

[20] Liang, K.J., Zhang, H.J., Liu, Y.Q., Zhang, Y., Wang, Y.Y. (2022). Classroom speech emotion recognition based on multi-channel convolution and SEnet network. In Artificial Intelligence in China, pp. 242–248. https://doi.org/10.1007/978-981-16-9423-3_30

[21] Li, M., Liu, M., Jiang, Z., et al. (2022). Multimodal emotion recognition and state analysis of classroom video and audio based on deep neural network. Journal of Interconnection Networks, 22(S4): 2146011. https://doi.org/10.1142/S0219265921460117

[22] He, J., Peng, L., Sun, B., Yu, L., Guo, M. (2021). Dual multi-task network with bridge-temporal-attention for student emotion recognition via classroom video. In 2021 International Joint Conference on Neural Networks (IJCNN), Shenzhen, China, pp. 1–8. https://doi.org/10.1109/IJCNN52387.2021.9533471

[23] Lei, L., He, Y., Li, X., Yu, S., Yin, Y., Qin, L., Liang, M. (2021). Classroom facial emotion recognition based on channel attent. In 2021 IEEE International Conference on Emergency Science and Information Technology (ICESIT), Chongqing, China, pp. 419–423. https://doi.org/10.1109/ICESIT53460.2021.9697038

[24] Liang, J., Zhao, X., Zhang, Z. (2020). Speech emotion recognition of teachers in classroom teaching. In 2020 Chinese Control and Decision Conference (CCDC), Hefei, China, pp. 5045–5050. https://doi.org/10.1109/CCDC49329.2020.9164823

8 Author

Suyun Wen, graduated from School of Foreign Studies, Hebei Normal University with a master's degree, is currently working in the Department of Public Foreign Languages, Shijiazhuang University of Applied Technology with a research direction of English education.

Article submitted 2022-11-03. Resubmitted 2022-12-09. Final acceptance 2022-12-11. Final version published as submitted by the authors.