Online-Journals.org

iJET International Journal of Emerging Technologies in Learning

iJET | eISSN: 1863-0383 | Vol. 18 No. 20 (2023) | OPEN ACCESS

https://doi.org/10.3991/ijet.v18i20.44219

PAPER

Establishing Teacher-Student Communicative Empathy in a Multimodal Interactive Environment and Skill Enhancement

Jinqiu Cai[1,2](✉)

[1]School of Foreign Studies, Suqian University, Suqian, China

[2]Faculty of Social Sciences and Liberal Arts, UCSI University, Kuala Lumpur, Malaysia

20204@squ.edu.cn

ABSTRACT

The multi-modal interactive environment is an important object of research in the field of modern education, and its influence on the communicative empathy of teachers and students gets increasingly obvious these days. However, the complex and dynamic features of communicative empathy of teachers and students are usually overlooked in existing studies, so this paper aims at investigating the teacher-student communicative empathy in a multi-modal interactive environment to fill in the said research gap. At first, the problem of feature fusion of communicative emotions of teachers and students in a multi-modal interactive environment was discussed, and a comprehensive and detailed model was established for measuring the features of these emotions. Then, problems of the recognition of emotional intentions of teachers and students and the establishment of their communicative empathy were studied, and a set of effective measures were proposed for enhancing the said communicative empathy, in the hopes of providing useful theoretical and practical evidences for noticing the importance of teacher-student communicative empathy in teaching and enhancing the ability of the said emotions.

KEYWORDS

multi-modal interactive environment, teacher-student communicative empathy, emotion feature fusion, emotional intention recognition, establishment of communicative empathy

1 INTRODUCTION

Information technology has brought fundamental changes to our living environment, and the multi-modal interactive environment has become an important part for the living, study, and work of modern people, owing to its merits of rich information expression methods and interaction modes [1–4]. In the field of education, multi-modal interactive environment gives teachers and students more convenient ways of communication as their communication is no longer limited to the

Cai, J. (2023). Establishing Teacher-Student Communicative Empathy in a Multimodal Interactive Environment and Skill Enhancement. *International Journal of Emerging Technologies in Learning (iJET)*, 18(20), pp. 183–198. https://doi.org/10.3991/ijet.v18i20.44219

Article submitted 2023-05-14. Revision uploaded 2023-08-04. Final acceptance 2023-08-15.

face-to-face communication, but can be conducted via multiple means including languages, body gestures, and facial expressions [5, 6]. However, despite the increasing application of multi-modal interactive environment and its positive influence on the communicative empathy of teachers and students, it hasn't received enough attention from field scholars yet [7–11].

In fact, researching the communicative empathy of teachers and students in a multi-modal interactive environment is of great theoretical and practical significance. One reason is that this research could help deepen our understanding of the features of emotions held by teachers and students during their communication process, so that the construction and improvement of teaching modes could be supported by solid theories, and teachers' teaching effect and students' learning experience could be both improved [12–14]. Another reason is that the research on emotional intention recognition and communicative empathy establishment can help us more accurately identify and understand the emotional needs of teachers and students during their communication process, thereby further smoothing their communication and ultimately improving teaching efficiency and quality [15, 16].

Although some scholars have noticed the research topic of teacher-student communicative empathy in the multi-modal interaction environment, there are a few defects with the existing works, such as some have neglected the complex and dynamic features of communicative empathy of teachers and students, and the research on emotion feature fusion and emotional intention recognition is far from sufficient [17, 18]. Besides, how to effectively establish teacher-student communicative empathy and enhance their ability of communicative empathy are questions pending for solutions, and this is also an urgent task for education practitioners.

To address the above issues, this paper attempts to explore the feature fusion of teacher-student communicative empathy, build a comprehensive and detailed model for measuring the said features and providing valuable references for teachers to detect, understand, and cope with students' emotional responses, and help them adjust their emotional expressions. Moreover, the recognition of emotional intentions of teachers and students and the establishment of their communicative empathy were investigated to discuss how to enhance their ability of communicative empathy via effective measures in specific teaching practice. This research provides new insights and methods for teaching practice, and is meaningful for promoting the progress of smart education.

2 FEATURE FUSION OF TEACHER-STUDENT COMMUNICATIVE EMOTIONS IN THE MULTI-MODAL INTERACTIVE ENVIRONMENT

In a multi-modal interaction environment, the recognition of emotional intentions of teachers and students is very important, especially for English teaching. The teaching process is not only about imparting knowledge, but also an exchange and sharing of emotions. For teachers, accurately identifying students' emotional intentions not only enables them to understand students' learning needs and difficulties, but also allows them to adjust the teaching strategies in a timely manner and enhance the teaching effect. For example, in English teaching, students' learning status is often exhibited through their emotions, such as when a student is trying hard to understand a difficult grammatical structure, he/she might show emotions of confusion or frustration, and by recognizing these emotional intentions, teachers

will be able to provide timely assistance to ease the learning anxiety of students and help them increase learning efficiency.

Out of these concerns, in this paper, a framework was constructed for teacher-student emotional intention recognition in a multi-modal interactive environment. The framework has three parts: data collection, data feature fusion, and emotion recognition. In the data collection stage, various types of data were collected so that the features of these data could be fused in the next stage. Data types that need to be collected include:

1. Speeches: speech data provide a lot of information about emotions and intentions, such as speech rate, pitch, and intonation, which can be captured through recording devices.
2. Texts: the content of teacher-student communications can be captured by recording dialogue texts or classroom notes, to provide information about emotional tendencies, context, and themes, etc.
3. Facial expressions: facial expressions are a very intuitive way of emotion expression, and can be captured in real time by cameras and their features can be attained by analyzing the recorded videos.
4. Body languages: body postures and hand gestures of teachers and students contain information about their emotions and intentions, and these data can be captured through video surveillance or body motion capture techniques.
5. Physiological conditions: data such as heart rate, blood pressure, and skin conductance of teachers and students, can be captured through physiological monitoring devices.

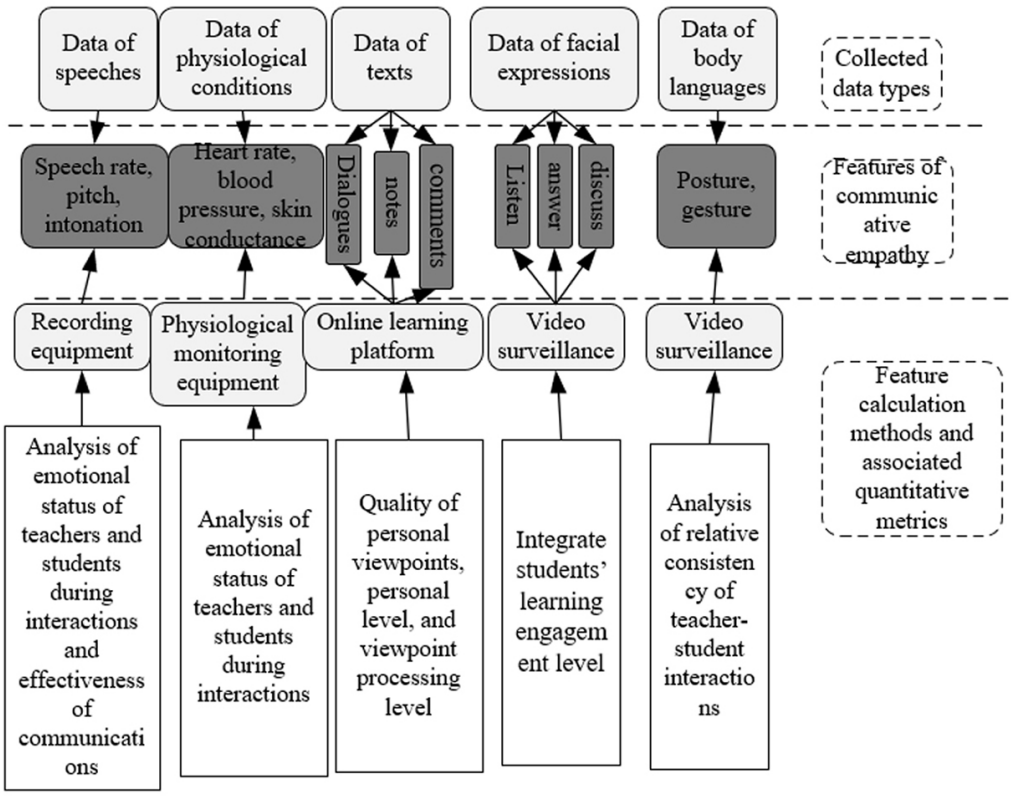


Fig. 1. Feature recognition method of collected data

All the above-mentioned data types have their respective collection methods and tools, some of which may require special equipment or software for acquisition and processing. The collected data will be processed further in the data feature fusion stage. Key features will be extracted to form a more representative high-dimensional dataset. Then, in the emotion recognition stage, these data will be fed into the RNNPB network, through the processing of which the emotional intentions of teachers and students will be recognized. Figure 1 shows the feature recognition method of collected data.

The expression of emotions during actual communications is not limited to a single mode (speeches or facial expressions), but is a collection of multiple modes. For instance, a person may express his/her emotions through multiple means including languages, facial expressions, and body language. The single-modal data may not be able to fully reflect a person's true emotions, while the fusion of multi-modal data can provide more comprehensive and accurate information about emotions. Here is an introduction of data feature fusion.

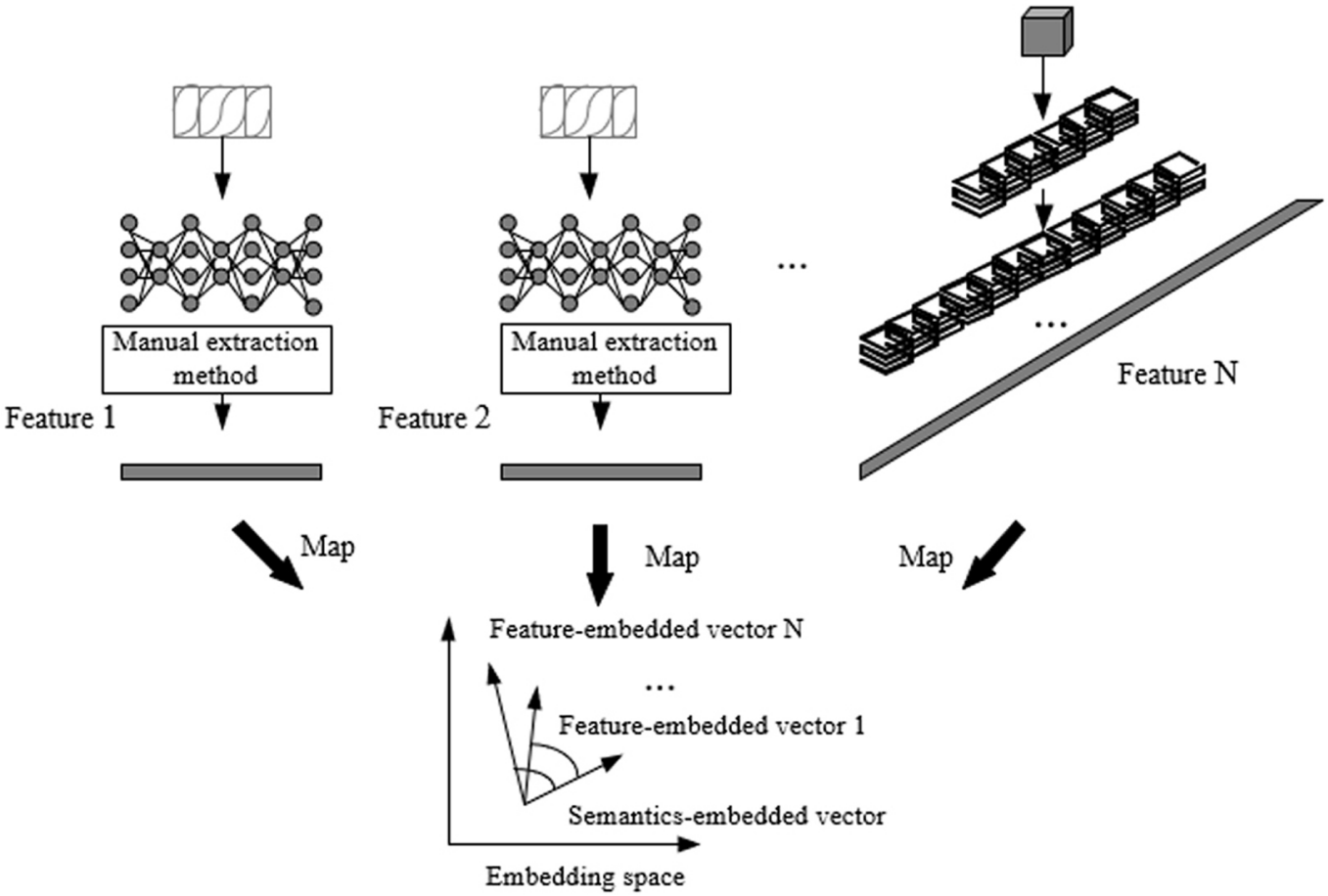


Fig. 2. Principle of emotion semantic analysis

Before emotional intention recognition, the collected data were subject to emotion semantic analysis, and the principle is given in Figure 2. The purpose of this process is to figure out the relations between each type of modal data and the emotions so as to facilitate the use of these data in the subsequent feature fusion stage. At first, the collected modal data were subject to feature encoding, and the purpose

of this step is to transform the raw and possibly unstructured data into a structured form for further processing and analysis. When conducting emotion semantic analysis, an emotion tag was set for each data sample. Valence and arousal are two commonly used emotion tags that respectively indicate the intensity and activity of emotions. These emotion tags can be regarded as the most direct association form of emotion semantics, which can help us understand and quantify the emotional information in the data. At last, the features and emotion tags were subject to correlation analysis, which is a key step for understanding information at the semantic level. In this paper, a *WSABIE*-based image annotation method was adopted, that is, the features and emotion tags were jointly mapped into a same potential embedding space. This method can help us understand the relations between features and tags, thereby providing a deeper-level understanding of the semantics of emotions.

Assuming: $a^{(u)}$ represents the extracted speech features, $b^{(u)}$ represents the extracted text features, $c^{(u)}$ represents the extracted facial expression features, $d^{(u)}$ represents the extracted body language features, $h_m^{(u)}$ represents each physiological condition feature, these five types of features were taken as the input; also, assuming: $x^{(u)}$ represents the tag corresponding to the feature set, E^F represents the shared embedding space to be mapped, $\Psi(c^{(u)})$ represents a linear mapping from feature space E^F, $\Psi\left(h_m^{(u)}\right)$ represents the linear transformation of physiological condition features, $\Theta(x^{(u)})$ represents the linear mapping from tags to the embedding space, then there are:

$$\Psi\left(a^{(u)}\right): E^f \to E^F \tag{1}$$

$$\Psi\left(b^{(u)}\right): E^f \to E^F \tag{2}$$

$$\Psi\left(c^{(u)}\right): E^f \to E^F \tag{3}$$

$$\Psi\left(d^{(u)}\right): E^f \to E^F \tag{4}$$

$$\Psi\left(h_m^{(u)}\right): E^{f_y} \to E^F \tag{5}$$

$$\Theta\left(x^{(u)}\right): \{1,2,\ldots,L\} \to E^F \tag{6}$$

Assuming: MO_m represents the m-th modal feature, by taking all collected data features and tags as input, a model could be established as follows:

$$d^{(u)} = \Theta\left(x^{(u)}\right)^Y \cdot \Psi\left(MO_m^{(u)}\right) \tag{7}$$

The similarity between features and tags can be measured by the size of $d^{(u)}$ value. Since $\Theta(\cdot)$ and $\Psi(\cdot)$ are linear transformations, let $Q_{x(u)}$ represent the $x^{(u)}$-th column of matrix Q, then there are:

$$\Theta\left(x^{(u)}\right) = Q_{x^{(u)}} \tag{8}$$

$$\Psi\left(MO_m^{(u)}\right) = C \cdot MO_m^{(u)} \tag{9}$$

When defining the objective function, it's generally hoped that the prior knowledge from the training set could be utilized. In the scenario of this study, objectives

of the model are to maximize the similarity between features and their corresponding tags and to minimize the similarity between features and non-corresponding tags, that is, the training model is expected to learn the correct feature-tag correspondences and avoid learning the wrong correspondences as much as possible. Based on these two objectives, the following objective function had been designed to guide the training process of the model and ultimately to achieve the goal of model performance improvement. Assuming: *U* represents the indicative function, then there are:

$$\underset{Q,C}{MIN}\sum_{u=1}^{L}\sum_{k\neq x^{(u)}}^{L}U\left(d_u\left(MO_m^{(u)}\right)\geq d^{(u)}\right) \tag{10}$$

$$d_u\left(MO_m^{(u)}\right)=Q_j^{YK}C\cdot MO_m^{(u)} \tag{11}$$

The training error of a single sample can be attained from the calculation of the following formulas:

$$LOSS_u=M\left(rank(u)\right) \tag{12}$$

$$rank(u)=\sum_{k\neq x^{(u)}}^{L}U\left(d_uMO_m^{(u)}+1\geq d^{(u)}\right) \tag{13}$$

To enable the model to share parameters in real time based on stochastic gradient descent, this paper introduced the idea of sampling into the *WARP* algorithm by defining a loss function as follows:

$$LOSS\left(d^{(u)},x^{(u)}\right)=\sum_{k\neq x^{(u)}}M\left(rank(u)\right)\frac{\left|1-d^{(u)}+d_k\left(MO_m^{(u)}\right)\right|_+}{rank(u)} \tag{14}$$

$$M(j)=\sum_{k=1}^{j}c_k \tag{15}$$

where, c_k is a constant.

The feature fusion method adopted in this study mainly uses the Logistic Regression (*LR*) and soft attention mechanism. The principle of feature fusion is shown in Figure 3. The values of *Q* and *C* could be attained through above calculation steps, and the mode-embedded vector could be attained based on the following formula:

$$O=C\cdot MO_m \tag{16}$$

The mode-embedded vector shown by the above formula was the closest to the semantic-embedded vector. For data of each mode, an *LR* model was trained, which takes the received semantic-embedded vector as the input. The output of *LR* is a probability value indicating the possibility of a particular tag. For each *LR* model, its training tag was set as:

$$t=\begin{cases}1,\ d^{(u)}>d_kMO_m^{(u)},k\neq x^{(u)}\\0,\ other\end{cases} \tag{17}$$

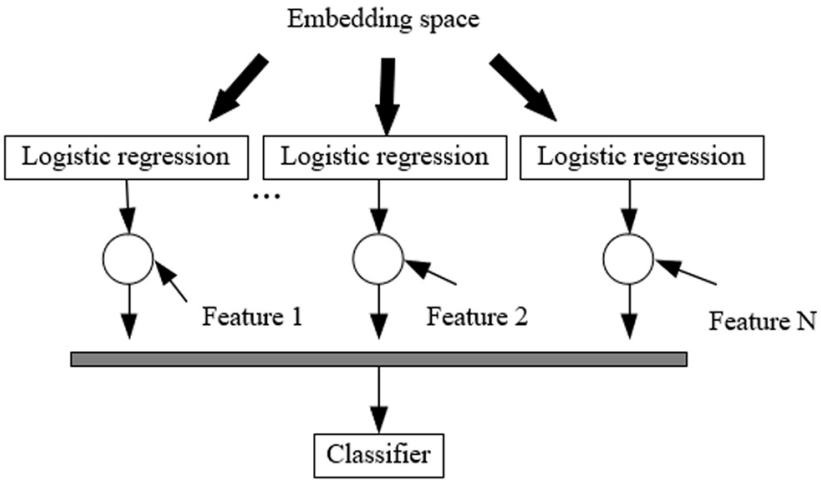


Fig. 3. Principle of feature fusion

Then, the probability values output by *LR* models were normalized and multiplied by the features of the corresponding mode; in this way, the features of each mode had attained a weight corresponding to its importance. In the meantime, each mode feature after subjected to weight adjustment was put in series and normalized together for once, and the attained result was taken as the input feature of the classifier. In the process of classifier training, instead of rigidly specifying the weight threshold of each mode, the proposed method flexibly adjusted the weight of each mode according to their respective contributions to the final classification result. After training, a group of hyper parameters was attained, and the corresponding probability can be obtained based on the following formula:

$$o = \frac{1}{1 + e^{-(s^Y z + n)}} \tag{18}$$

With the help of the soft attention mechanism, the feature fusion method proposed in this paper can effectively avoid the influence of outliers in case of rigid specification and enhance the robustness of the model. By flexibly adjusting the weight of each mode during classifier training, the method can better mine the effective information of each mode instead of artificially specifying the weights. Through feature weight adjustment and the soft attention mechanism, the information of different modes could be fused better and the prediction accuracy of the model could be improved. Overall, the method combines the merits of *LR* and soft attention mechanism, and it can effectively process the multi-modal data and improve the performance of the model.

3 TEACHER-STUDENT EMOTIONAL INTENTION RECOGNITION AND COMMUNICATIVE EMPATHY ESTABLISHMENT IN THE MULTIMODAL INTERACTIVE ENVIRONMENT

The *RNNPB* network is a recurrent neural network capable of processing sequential data, and this quality allows the network to well handle data that change continuously in the time dimension, such as speeches, texts, and body languages, and it is very effective for understanding and capturing changes in emotions during the continuous communications. Compared with conventional *RNN*, the *RNNPB* network is

better at capturing long-term dependencies when dealing with long sequential data, which is conductive to understanding changes in emotions and intentions during communication over a long time scale. These merits of *RNNPB* make it a powerful tool for dealing with the task of recognizing the emotional intentions of teachers and students in the multi-modal interactive environment, and Figure 4 illustrates the network structure.

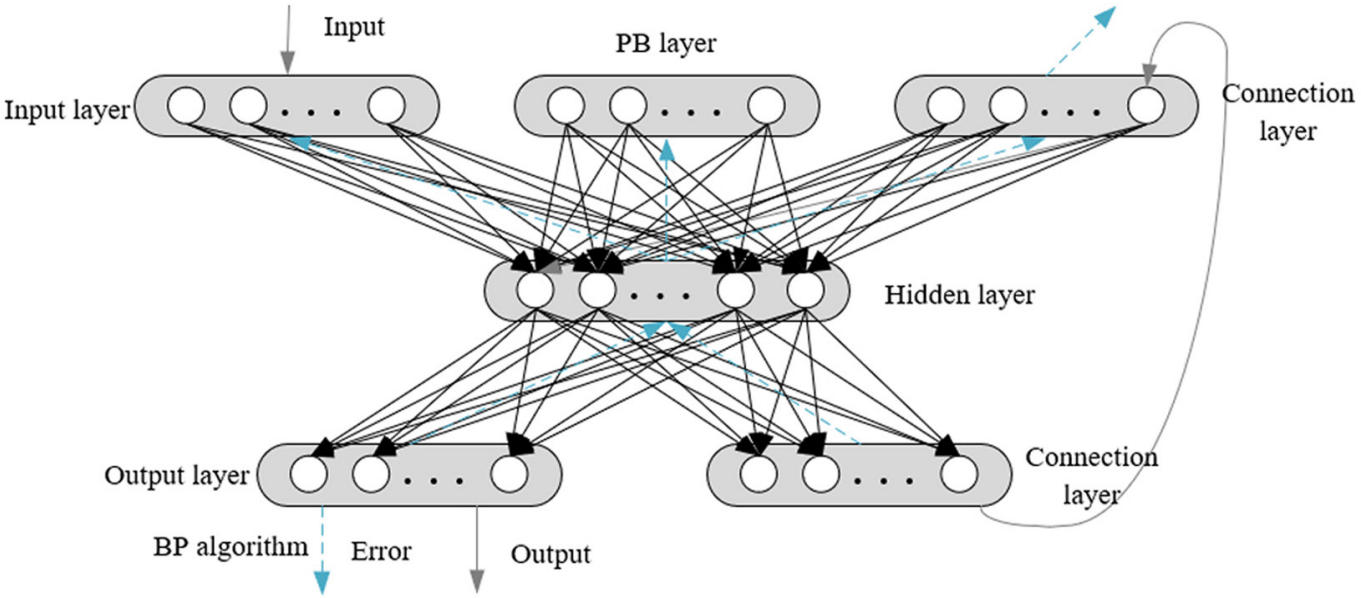


Fig. 4. Structure of the *RNNPB* network

The *RNNPB* network is mainly composed of 5 network layers: input layer, hidden layer, *PB* layer, connection layer, and output layer. Assuming: q_{gu} represents the weight between nodes in the hidden layer and input layer, q_{pg} represents the weight between nodes in the hidden layer and output layer, q_{go} represents the weight between nodes in the *PB* layer and hidden layer, and q_{gv} represents the weight between nodes in the connection layer and hidden layer; $h_u(y)$ and $h_g(y–1)$ represent activation functions, subscripts u and g represent parameters of the input layer and hidden layer, $PB_j(y)$ represents the activation function of the *PB* layer, y represents the time step, then the combination of the outputs of input layer, *PB* layer and connection layer forms $T_g(y)$, the input of the hidden layer, that is:

$$t_g(y) = \sum_u h_u(y) q_{gu} + \sum_v h_u(y-1) q_{gv} \sum_j PB_j(y) q_{go} \tag{19}$$

All activation functions in the network had adopted the *sigmoid* function, and there are:

$$sigmoid(y) = 1.716 \cdot \tanh\left(\frac{2}{3} y\right) \tag{20}$$

$$PB_j = sigmoid\left(\frac{2}{3} \vartheta_j\right) \tag{21}$$

Assuming: r represents the number of times the entire training set is traversed by the training algorithm, ϑ_j represents the output of the *PB* layer, $\sigma_{j,y}^{ON}$ represents

the back propagation error of the *PB* layer at time step *y*, ε_j represents the learning rate of the *PB* layer, then the update of the internal value of the *j*-th PB unit can be calculated by the following formula:

$$\vartheta_j(r+1) = \vartheta_j(r) + \varepsilon_j \sum_{y=1}^{Y} \sigma_{j,y}^{ON} \tag{22}$$

The relationship between ε_j and $\sigma_{j,y}^{ON}$ can be expressed as:

$$\varepsilon_j \propto \frac{1}{Y} \left\| \sum_{y=1}^{Y} \sigma_{j,y}^{ON} \right\| \tag{23}$$

Assuming: $gh_j^f(d+1)$ represents the expected output, $h_j^p(y)$ represents the actual output, *Y* represents the length of each sense-motion timer series, *B* represents the number of nodes in the output layer, then the cost function of the training dataset can be expressed as:

$$K = \frac{1}{2} \sum_{y}^{Yy} \sum_{j}^{B} \left(h_j^f(y+1) - h_j^p(y) \right)^2 \tag{24}$$

The weight update of the *RNNPB* network conforms to the law of gradient descent, then there is:

$$\Delta q_{uk} = -\varepsilon_{uk} \frac{\partial K}{\partial q_{uk}} \tag{25}$$

The calculation of learning rate needs to consider the variation of weight q_{uk} between neurons *u* and *k* in two consecutive training periods, then there is:

$$\gamma_{uk} = \frac{\partial K}{\partial q_{uk}}(y-1) \cdot \frac{\partial K}{\partial q_{uk}}(y) \tag{26}$$

Let the acceleration or deceleration of learning rate be determined by $\xi^+>1$ or $\xi^-<1$. The maximum and minimum values of learning rate ε_{uk} are represented by ε_{MAX} and ε_{MIN}, respectively, then the update rule of learning rate is given by the following formula:

$$\varepsilon_{uk}(r) = \begin{cases} MAX\left(\varepsilon_{uk}(r) \cdot \xi^-, \varepsilon_{MIN}\right); \text{ or } \gamma_{uk} > 0 \\ MIN\left(\varepsilon_{uk}(r) \cdot \xi^+, \varepsilon_{MAX}\right); \text{ or } \gamma_{uk} > 0 \\ \varepsilon_{uk}(r-1); \ others \end{cases} \tag{27}$$

After completing the teacher-student emotional intention recognition in the multimodal interactive environment, measures should be taken to further establish communicative empathy between teachers and students and enhancing their ability. At first, the recognition results will be fed back to the teachers, which can be accomplished through real-time panels or regular reports, and these allow the teachers to better understand students' emotional status and their intentions. The sharing of such information is helpful to establish empathy between teachers and students, since it enables teachers to better understand students' needs and moods. Teachers should develop appropriate emotional response strategies based on the recognition results. For example, if a student is identified as anxious or upset, the teacher

may respond to the student's emotions by giving encouragement or reassurance. The development and implementation of such strategies can help improve the effect and efficiency of communications between teachers and students.

Next, teachers can use the recognition results to improve their communication methods and strategies. For example, if a teacher finds that his/her teaching style may confuse the students or make them anxious, then the teacher can seek further training or support to improve his/her communication skills. By analyzing the results of emotional intention recognition, teachers can optimize the multi-modal interaction environment as well. For example, if a certain mode (such as video or audio) is found to be more effective in conveying emotional messages, teachers can give this mode more considerations in future teaching design.

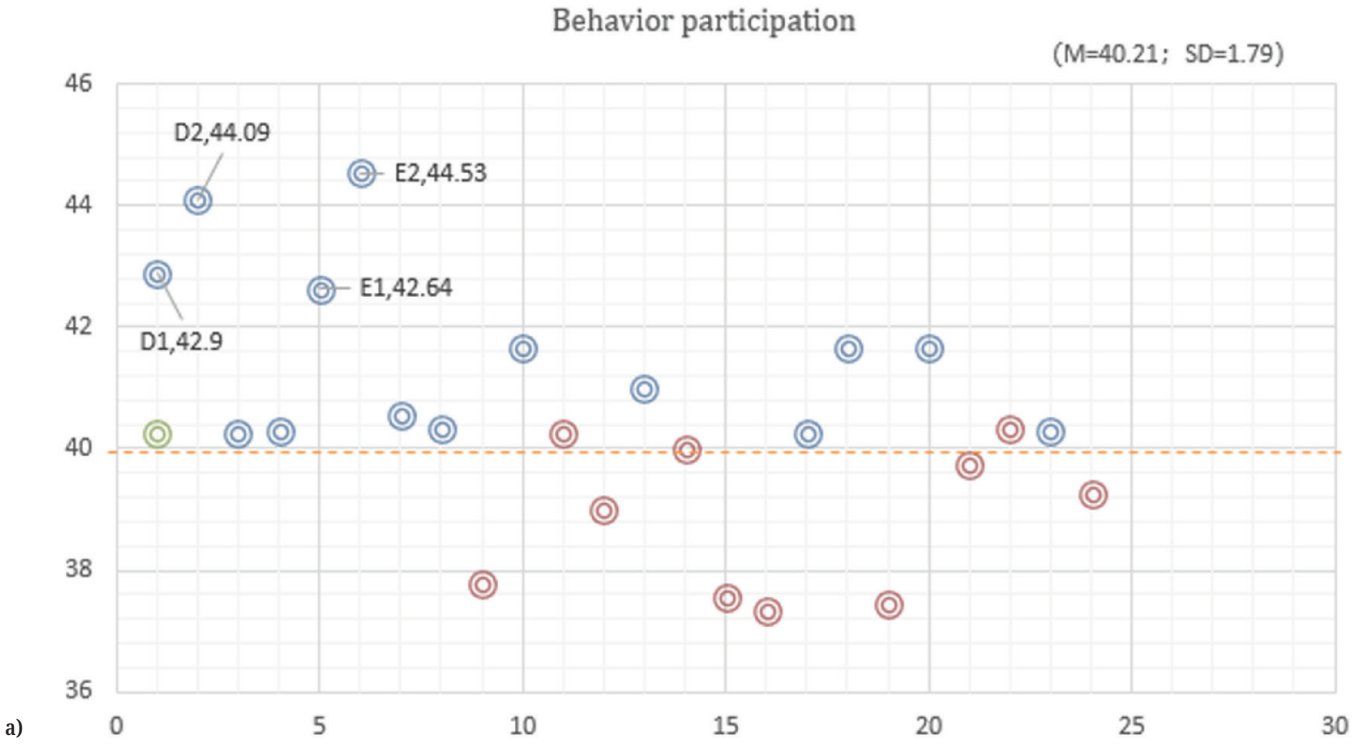
The above steps constitute a cyclical process that not only promotes the establishment of communicative empathy between teachers and students, but also enhances their communication ability constantly.

4 EXPERIMENTAL RESULTS AND ANALYSIS

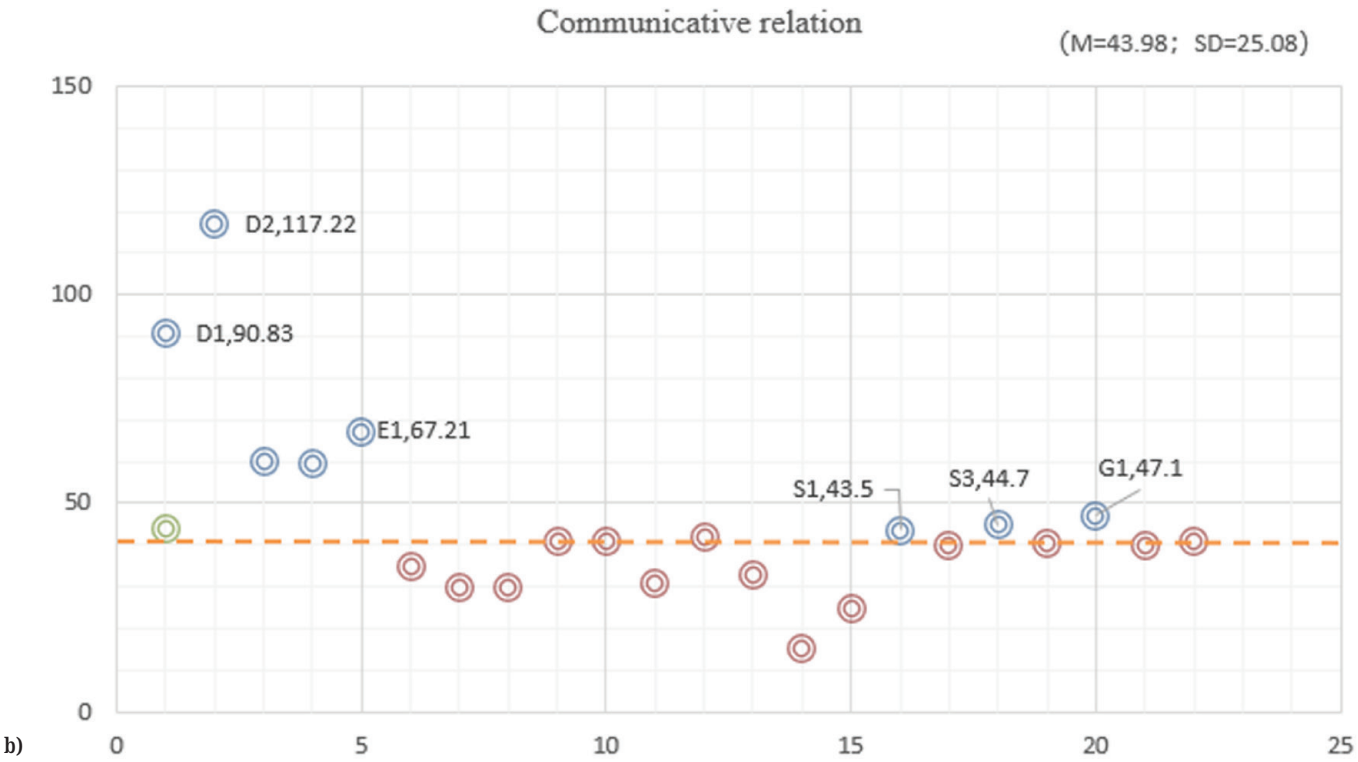
The valence and arousal recognition results of the features of five types of extracted data could be analyzed based on the data given in Table 1. In terms of the recognition of valence, the accuracy of texts reached 56.2%, the performance was the best, followed by physiological conditions with an accuracy of 54.8%. The accuracy of speeches and facial expressions was 52.6% and 51.7%, respectively. Both were low, but their F1 values reached 36.4% and 54.1%, and the performance was good, indicating that in terms of the recognition of valence, the proposed feature extraction method can get ideal results and the performance of each mode was relatively average. In term of the recognition of arousal, the accuracy of facial expressions and physiological conditions both led the way with a value of 52.8%. Although the accuracy of speeches and body languages was low, their UAR values respectively reached 62.1% and 53.1%, indicating that when recognizing the arousal, the performance on the two types of extracted data was relatively good. Overall speaking, the *DBN*-based method that combines manual extraction had achieved ideal effect in recognizing teacher-student emotional intentions in the multimodal interactive environment. Each type of mode data had some contributions to the recognition result of emotional intentions, which has verified the effectiveness of adopting multimodal data to conduct emotional intention recognition. In the meantime, the effect of the proposed method also indicates that appropriate feature extraction and processing techniques can greatly increase the utilization efficiency of multimodal data and improve the accuracy of emotional intention recognition.

Table 1. Recognition results of valence and arousal of 5 types of extracted data features

		Speeches	Texts	Facial Expressions	Body Languages	Physiological Conditions
Valence	Accuracy	52.6	56.2	51.7	52.3	54.8
	UAR	51.0	55.1	52.3	51.2	52.3
	*F*1 value	36.4	37.3	54.1	36.5	36.8
Arousal	Accuracy	47.8	51.5	52.8	43.8	52.8
	UAR	62.1	36.8	52.4	53.1	21.2
	*F*1 value	52.6	33.6	44.3	39.4	2.1



a)



b)

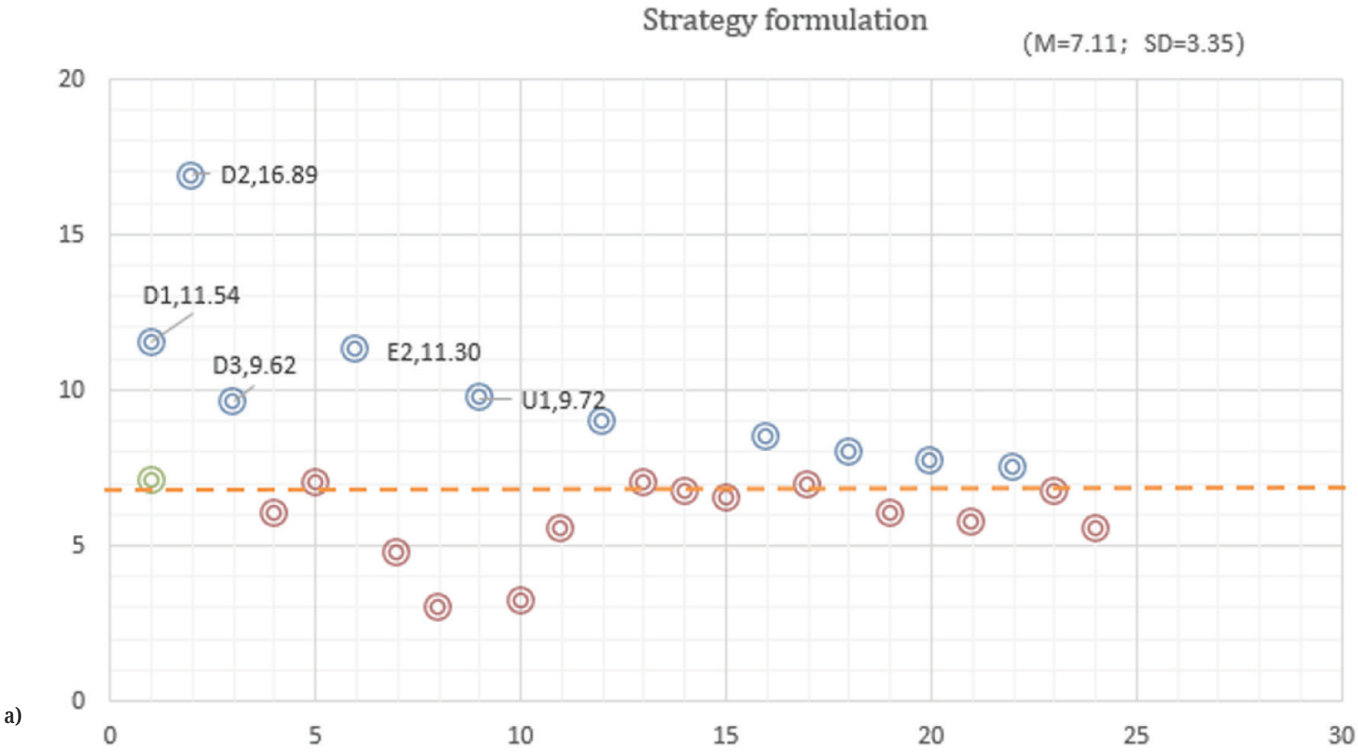
Fig. 5. Basic situations of communicative empathy establishment and ability enhancement of students

The situations of communicative empathy establishment and ability enhancement of students were discussed from two aspects of behavior participation and communicative relation. Figure 5a shows that in the behavior participation

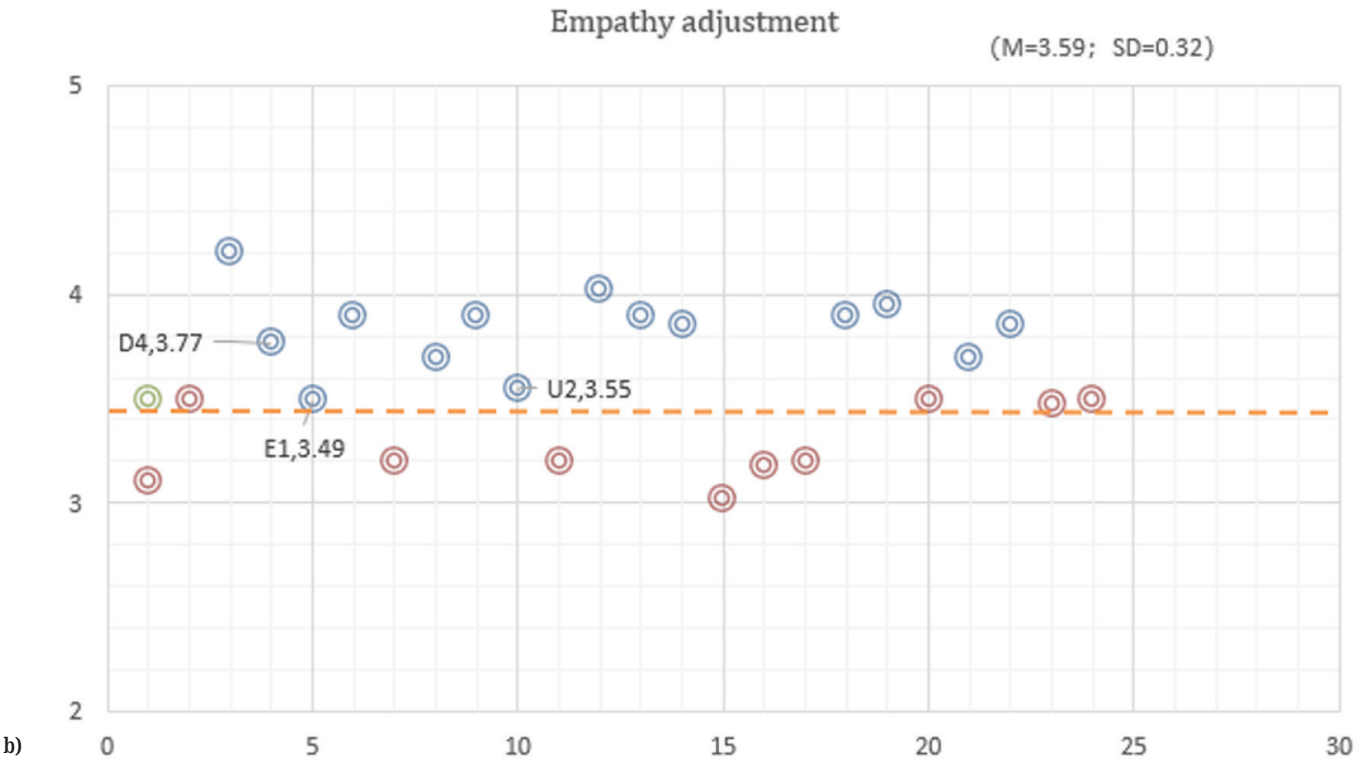
dimension, the difference in students' ability of communicative empathy was small (standard deviation = 1.79). For example, the degree of engagement in interactions of students D1, D2, E1, E2 was significantly higher than that of other students, but for most students, the ability level of communicative empathy was concentrated around the line of mean value (M=40.21), indicating that the majority of students were similar in terms of the attention degree and participation degree of multimodal interaction tasks. Figure 5b shows that in the communicative relation dimension, the difference in students' ability of communicative empathy was relatively large (standard deviation = 25.8). For example, students D1, D2, and E1 were much more competent than others in maintaining and establishing a communicative relation in which teachers and students rely on each other. Meanwhile, some students such as S1, S3, and G1 had an ability level of communicative empathy close to the mean value (M = 43.98), which implies that there was a large difference among students in establishing and maintaining the interactive relationship between teachers and students. Analysis of the two dimensions shows that in a multimodal interaction environment, for most students, their attention degree and participation degree of interaction tasks were consistent, but there's a large difference in the ability to maintain and establish the interactive relationship between teachers and students.

The situations of communicative empathy establishment and ability enhancement of students were discussed from two aspects of communication strategy formulation and empathy adjustment. Figure 6a shows that in the dimension of strategy formulation, the difference in teachers' ability level of communicative empathy was large (standard deviation = 3.35). Teachers D1, D2, D3, E2, and Ut were significantly more competent than others in terms of the ability of communicative empathy, but for most teachers, their ability level of communicative empathy was lower than the mean value (M = 7.11), indicating that there's a large difference in the efforts made by teachers in establishing the empathy, such as the process of knowledge sharing and negotiation. Figure 6b shows that in the dimension of empathy adjustment, the difference in teachers' ability level of communicative empathy was small (standard deviation = 0.52). For most teachers, the ability level fluctuated around the mean value (M = 3.59), such as teachers D4, E1, and U2, and this result indicates that the teachers were basically consistent in the degree of planning, monitoring, and reflecting the multi-modal interaction tasks. In summary, teachers showed significant differences in the ability of empathy in the aspect of communication strategy formulation, but they were relatively consistent in the aspect of empathy adjustment.

Table 2 gives statistics of four dimensions of behavior participation, communicative relation, strategy formulation, and empathy adjustment, and lists the mean, standard deviation, upper limit, and lower limit of each dimension. These data can help us have a deeper understanding of the distribution and change range of teacher-student communicative empathy ability. According to the data in the table, in terms of behavior participation and empathy adjustment, the difference in the ability of communicative empathy of teachers and students was small, indicating that they were consistent in empathy establishment and ability enhancement in these two dimensions. In contrast, in dimensions of communicative relation and strategy formulation, they exhibited significant difference in the ability of empathy, and this reminds us that these differences should be investigated and analyzed in future research and teaching practice so as to better enhance the communicative empathy ability of teachers and students.



a)



b)

Fig. 6. Basic situation of communicative empathy establishment and ability enhancement of teachers

Table 2. Division of teacher-student communicative empathy abilities

Dimension	Mean	Standard Deviation	Upper Limit	Lower Limit
Behavior participation	41.25	1.98	41.58	37.87
Communicative relation	42.68	23.26	65.39	18.39
Strategy formulation	7.12	3.42	11.58	3.56
Empathy adjustment	3.58	0.31	3.89	3.45

Figure 7 shows the distribution of the ability levels of respondents in terms of four dimensions of communicative empathy (behavior participation, communicative relation, strategy formulation, and empathy adjustment), and the ability was divided into four levels in this study: low-level, mid-level, high-level, and super high-level. As can be known from the figure, for most respondents who participated in the survey, their ability of communicative empathy was at the mid-level, which may indicate that they can well handle most situations in daily communications. Some of them showed obvious advantages in the ability of some dimensions, especially in terms of communicative relation and empathy adjustment. Many people had a super high-level ability, and this may indicates that these people are good at understanding other people's emotions and needs and they can establish and maintain good interpersonal relationships. While for those who are weaker in the ability of each dimension, they might need more instructions and training to enhance their ability of communicative empathy.

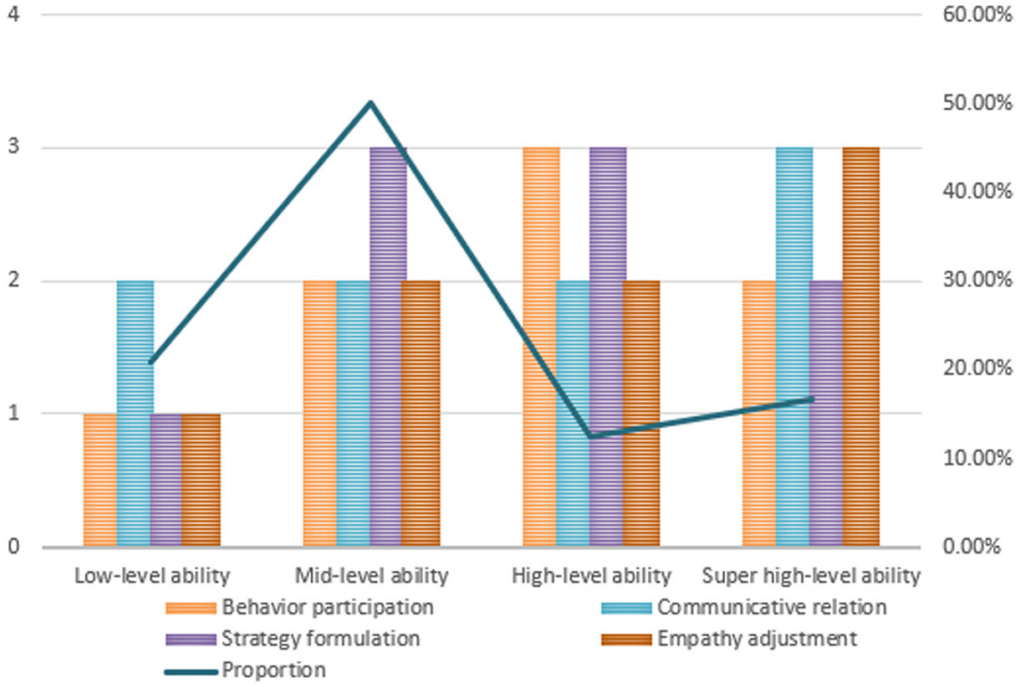


Fig. 7. 4 Clustered groups of communicative empathy

5 CONCLUSION

This study investigated the problem of emotional intention recognition of teachers and students via a multi-modal (data of speeches, texts, facial expressions, body languages, and physiological conditions) emotion recognition method. Through emotion semantic analysis of the collected data, the features and emotion tags were

mapped into a potential embedding space, and an emotion recognition model was built, based on which the objective function was defined and optimized, the similarity of training set features and tags was maximized, and the recognition accuracy of the model was improved.

Through the fusion of multimodal features, this paper adopted a classifier model constructed based on LR and introduced the soft attention mechanism to more flexibly adjust the weight of different modes, thereby enhancing the generalization ability of the model. For the evaluation of the effect of five data collection methods, a method that combines *DBN* and manual extraction was adopted and the extracted features could get an ideal recognition effect in most cases.

On this basis, this paper gave an in-depth study on the communicative empathy establishment and ability enhancement of teachers and students in the multimodal interactive environment, and the results suggest a small difference in students' behavior participation and communicative relation, and a large difference in the establishment of communicative relation. As for teachers, they vary greatly in the ability of strategy formulation but there's not much difference in the ability of empathy adjustment. These findings have important implications for understanding and enhancing the communicative empathy of teachers and students.

In summary, this study improved the accuracy and generalization ability of the emotion recognition model through emotion analysis and multimodal feature fusion techniques, and provided new theoretical and practical evidences for the communicative empathy establishment and ability enhancement of teachers and students. However, it should be noted that there are large differences in the ability of communicative empathy between teachers and students, and targeted training and instructions are needed to further improve their communication effect.

6 REFERENCES

[1] N. T. H. Giang, P. T. T. Hai, N. T. T. Tu, and P. X. Tan, "Exploring the readiness for digital transformation in a higher education institution towards industrial revolution 4.0," *International Journal of Engineering Pedagogy,* vol. 11, no. 2, pp. 4–24, 2021. https://doi.org/10.3991/ijep.v11i2.17515

[2] F. M. Zain, S. N. Sailin, M. Kasim, A. M. Abdul Karim, and N. N. Zameri, "Developing an augmented reality immersive learning design (AILeaD) framework: A fuzzy delphi approach," *International Journal of Interactive Mobile Technologies,* vol. 16, no. 11, pp. 65–90, 2022. https://doi.org/10.3991/ijim.v16i11.30063

[3] V. Amnouychokanant, S. Boonlue, S. Chuathong, and K. Thamwipat, "Online learning using block-based programming to foster computational thinking abilities during the COVID-19 pandemic," *International Journal of Emerging Technologies in Learning,* vol. 16, no. 13, pp. 227–247, 2021. https://doi.org/10.3991/ijet.v16i13.22591

[4] I. Holik, T. Kersánszki, G. Molnár, and I. D. Sanda, "Teachers' digital skills and methodological characteristics of online education," *International Journal of Engineering Pedagogy,* vol. 13, no. 4, pp. 50–65, 2023. https://doi.org/10.3991/ijep.v13i4.37077

[5] M. F. Ramadhan, M. Mundilarto, A. Ariswan, I. Irwanto, B. Bahtiar, and S. Gummah, "The effect of interface instrumentation experiments-supported blended learning on students' critical thinking skills and academic achievement," *International Journal of Interactive Mobile Technologies,* vol. 17, no. 14, pp. 101–125. https://doi.org/10.3991/ijim.v17i14.38611

[6] N. Khamcharoen, T. Kantathanawat, and A. Sukkamart, "Developing student Creative Problem-Solving Skills (CPSS) using online digital storytelling: A training course development method," *International Journal of Emerging Technologies in Learning,* vol. 17, no. 11, pp. 17–34. https://doi.org/10.3991/ijet.v17i11.29931

[7] R. Hu and X. Wang, "Multimodal teaching strategy of college English based on computer technology," in *2021 4th International Conference on Information Systems and Computer Aided Education,* 2021, pp. 227–230. https://doi.org/10.1145/3482632.3482679

[8] P. Lameras, S. Philippe, and L. Oertel, "A serious game for amplifying awareness on multimodal teaching: Game design and usability study," in *Internet of Things, Infrastructures and Mobile Applications: Proceedings of the 13th IMCL Conference* 13, 2021, pp. 559–570. https://doi.org/10.1007/978-3-030-49932-7_53

[9] D. Weber, W. Fuhl, E. Kasneci, and A. Zell, "Multiperspective teaching of unknown objects via shared-gaze-based multimodal human-robot interaction," in *HRI 2023-Proceedings of the 2023 ACM/IEEE International Conference on Human-Robot Interaction,* 2023, pp. 544–553. https://doi.org/10.1145/3568162.3578627

[10] S. Philippe, A. D. Souchet, P. Lameras, P. Petridis, J. Caporal, G. Coldeboeuf, and H. Duzan, "Multimodal teaching, learning and training in virtual reality: A review and case study," *Virtual Reality & Intelligent Hardware,* vol. 2, no. 5, pp. 421–442, 2020. https://doi.org/10.1016/j.vrih.2020.07.008

[11] Y. Pan, J. Wu, R. Ju, Z. Zhou, J. Gu, S. Zeng, L. Yuan, and M. Li, "A multimodal framework for automated teaching quality assessment of one-to-many online instruction videos," in *2022 26th International Conference on Pattern Recognition (ICPR),* Montreal, QC, Canada, pp. 1777–1783, 2022. https://doi.org/10.1109/ICPR56361.2022.9956185

[12] Y. Chen, L. Lin, and X. Wang, "Online teacher-student interpersonal management under crisis conditions: A sentiment analysis through machine learning," in *2022 International Conference on Cloud Computing, Big Data and Internet of Things (3CBIT),* Wuhan, China, pp. 48–52, 2022. https://doi.org/10.1109/3CBIT57391.2022.00019

[13] Y. Zhang, "Influence of teacher-student interaction on course learning effect in distance education," *International Journal of Emerging Technologies in Learning,* vol. 17, no. 10, pp. 215–226. https://doi.org/10.3991/ijet.v17i10.30913

[14] Y. Liu and W. Qi, "Application of flipped classroom in the era of big data: What factors influence the effect of teacher-student interaction in oral English teaching," *Wireless Communications and Mobile Computing,* vol. 2021, pp. 1–7, 2021. https://doi.org/10.1155/2021/4966974

[15] H. Song, J. Kim, and W. Luo, "Teacher–student relationship in online classes: A role of teacher self-disclosure," *Computers in Human Behavior,* vol. 54, pp. 436–443. https://doi.org/10.1016/j.chb.2015.07.037

[16] D. Xin and Y. Shi, "Modeling and analysis of UML-based teacher-student intercommunion platform," in *2009 Second International Conference on Intelligent Computation Technology and Automation,* Changsha, China, pp. 90–93. https://doi.org/10.1109/ICICTA.2009.259

[17] S. Jothimani and K. Premalatha, "THFN: Emotional health recognition of elderly people using a Two-Step Hybrid feature fusion network along with Monte-Carlo dropout," *Biomedical Signal Processing and Control,* vol. 86, pp. 105116, 2023. https://doi.org/10.1016/j.bspc.2023.105116

[18] I. Zubiaga and R. Justo, "Multimodal feature evaluation and fusion for emotional well-being monitorization," in *Iberian Conference on Pattern Recognition and Image Analysis,* pp. 242–254, 2022. https://doi.org/10.1007/978-3-031-04881-4_20

7 AUTHOR

Jinqiu Cai is currently a PhD candidate in Education at USCI University, Malaysia, and is working within the School of Foreign Studies at Suqian University, China. Her research interests encompass English education, British and American literature, and translation studies. She has published more than three research papers and produced two translation works (E-mail: 20204@squ.edu.cn; ORCID: https://orcid.org/0000-0002-6266-7471).