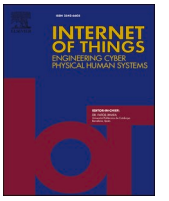
Internet of Things 25 (2024) 101069

Contents lists available at ScienceDirect

Internet of Things

journal homepage: www.sciencedirect.com/journal/internet-of-things



A multimodal teacher speech emotion recognition method in the smart classroom

Gang Zhao, Yinan Zhang *, Jie Chu

School of Educational Information Technology, Faculty of Artificial Intelligence Education, Central China Normal University, Wuhan, China

ARTICLE INFO

Keywords:
Teacher emotion
Smart classroom
Multimodal emotion recognition
Prosodic feature

ABSTRACT

As a collection of various IoT devices, smart classroom can record various forms of teaching data and provide rich data for recognizing teachers' emotions. Recognizing and analyzing teachers' emotions can promote teachers' professional development. Nowadays, most of the automatic emotion recognition methods for teachers in smart classroom are based on facial expressions. However, since teachers usually keep smiling to mobilize the classroom atmosphere, the recognition results may not reflect the real mental state of teachers. By observing teaching videos, it is found that the prosody and text in the teachers' speech can reflect the implicit emotion of the teacher. Therefore, a multimodal teacher emotion dataset (MTED) was built based on teaching videos recorded by IoT cameras and microphones in smart classroom. A neural network combining multiple prosodic features and text content for teacher speech emotion recognition is proposed. The proposed method fills the gap in teacher speech emotion recognition, our proposed method has higher accuracy. Experimental results show that ProsodyBERT achieves 78.6 % UA_4 and 66.2 % UA_6 on IEMOCAP and MELD, respectively, surpassing the existing methods. The proposed method reached 82.1% UA_6 on MTED self-built dataset, which is 9.6 %-21.4 % higher than that of unimodal method in teacher emotion recognition. An ablation experiment is designed and implemented on MTED dataset to explore the influence of each module in ProsodyBERT on teacher speech emotion recognition task. The experimental results in the smart classroom record show that ProsodyBERT has higher accuracy and stronger robustness than unimodal methods.

Introduction

With the application and popularization of the Internet of Things (IoT), the IoT-based smart classroom has gradually become the main application scenes of teaching analysis. The smart classroom is constructed by IoT devices on the basis of interconnection, these devices combine physical space with digital space, can provide personalized support services for learning. With the support of IoT, the IoT devices such as cameras and microphones embedded in smart classrooms can record massive of audio and video data. These data imply the emotion, behavior, learning status and other information of teachers and students. Among them, emotion has high information value, it can reflect a teacher's mental state and attitude [1]. Recognizing and analyzing teachers' emotion is conducive to capturing teachers' current mental state, self-approval and other information to a certain extent. This work can provide powerful data support for teachers to optimize teaching utterance and reflect on teaching process, enhance teachers' emotional expression ability, and thus guarantee teachers' professional development [2]. For massive data recoded by IoT devices, accurate and automatic recognition methods are needed. Therefore, how to efficiently recognize teacher's emotions in smart classroom has attracted much

* Corresponding author.

https://doi.org/10.1016/j.iot.2024.101069

Available online 13 January 2024

attention. Educational researchers still use the manual labeling method to measure teacher emotions at present. This method relies on subjective evaluation and manual annotation of teachers' emotions, which is easy to cause problems such as inconsistent evaluation standards, time-consuming and laborious.

Nowadays, deep-learning based speech emotion recognition (SER) is applied to teaching evaluation, which improves the teaching quality through real-time analysis and timely feedback [3]. In the domain of emotion recognition, the prevailing approach involves the extraction of emotional features from various sources such as facial expressions, body movements, text content, speech waveforms, and other relevant attributes. Due to the teacher's facial expressions are very intuitive and easy to capture, the existing researches tend to reflect the teacher emotion by recognizing the teacher's facial expressions. However, the teachers tend to liven up the class by smiling [4]. Therefore, recognizing teacher emotion only from expression may cause the gap between the recognition result and the real emotion state. From the perspective of utterance, teachers will unconsciously imply their emotional state when they speak. Thus, exploring teacher emotion recognition methods from the perspective of utterance seems to be more consistent with the real emotions. At present, there are deep- learning based teacher emotion recognition methods, but the existing methods only consider the teacher emotion from a single modality, and the effect is not significant. Dong [5] has indicated that the efficiency of emotion recognition can be effectively improved by integrating multi-source emotion features.

Through observing massive teaching videos in smart classrooms, it is found that both prosodic features and text content can reflect teacher's speech emotion, but their contributions are not consistent in different emotional types. Besides, not all speech frames have the same contribution to emotion recognition, and some key frames can better reflect the emotional characteristics of the teacher. According to these findings, this paper proposed a ProsodyBERT neural network that can utilize multiple prosodic features and multimodal information. ProsodyBERT is composed of a prosody encoder, a text encoder and a feature decoder, which has ability to combine text features and prosodic features. By introducing an attention pooling layer, the prosody encoder can capture the utterance-level features and frame-level features of prosody at the same time. The proposed method was evaluated on the public dataset, and the experimental results prove the superiority of the proposed method to a certain extent, which surpasses the current state-of-the-art methods. And the proposed method is more robust than the traditional methods in the smart classroom. The main contributions are as follows:

(1) By observing lots of teaching videos, we found and summarized the prosodic characteristics of emotion in teacher utterance. According to the characteristics we found, a teacher speech emotion recognition (TSER) method combining prosodic features and text features is proposed, which makes full use of multimodal auditory data in intelligent classrooms and makes up for the shortcomings of current emotion recognition methods in teaching scenes.

(2) With the information advantage provided by the smart classroom, a neural network ProsodyBERT was constructed. ProsodyBERT can integrate prosodic features and textual features of teacher utterance and select key speech frames through attention pooling mechanism. ProsodyBERT performs very well in the smart classroom.

The rest of the paper is organized as follows: In Section 2, we analyzed the limitations of existing work on teacher emotion recognition, as well as the current related work on multimodal emotion recognition. In Section 3, the proposed TSER method that can integrate prosodic features and text content is introduced in detail. Experimental results and discussion of our proposed method are presented in Section 4, and conclusions are given in Section 5.

Related work

Emotion recognition in teaching scenes

Teacher's emotion is an important reference for analyzing teacher's teaching ability. Anttila [6] explored the emotion of 19 trainee teachers in different academic activities, recorded and reported various types of teacher emotions, and provided data support for the professional development of trainee teachers. Frenzel [7] analyzed the influence of teacher emotion on student outcome by constructing the framework of teacher emotion and student outcome. Before analyzing teacher emotion, recognizing teachers' emotion is a necessary prerequisite. Teacher emotion can be expressed in multiple dimensions, such as speech state, facial activity, etc. With the application of IoT-based devices, the IoT devices in smart classroom provide sufficient conditions to empower emotion recognition. In addition, the rapid development of artificial intelligence technology based on deep learning in recent years has created opportunities for the integration of artificial intelligence and IoT. Based on this, researchers have proposed different methods for emotion recognition from different dimensions. Dukić [8] et al. constructed a real-time emotion prediction method for video analysis through Inception-v3 and ResNet34, which was used to analyze the emotional state of students in videos. Gao [9] used a web camera to collect teachers' facial expressions in classroom, designed and proposed an emotion recognition method based on teachers' facial expressions for recognizing the emotion of political teachers. Zhu [10] et al. explored a facial-based approach to student emotion recognition for learning scenes supported by smart classrooms. Fakhar [11] et al. developed a real-time automatic emotion recognition system to obtain real-time input from the cameras deployed in smart classroom, thus enabling the emotion monitoring in the smart classroom. It can be seen that the emotion recognition in smart classroom is still mainly based on facial expression recognition. However, teachers are often required to maintain a positive posture to enliven the positive learning atmosphere in the classroom and maintain a smiling state for a long time [12]. Therefore, analyzing teacher emotions only by recognizing their facial expressions is hardly applicable to real classrooms. Some studies focus on the emotional states implied in the utterances of teachers and students, such as acoustic features and utterance characteristics. For example, based on the convolutional neural network, Wang [13] used audio equipment to extract the artificial features of spectrogram and Mel-spectrogram from students' speech, and realized student speech emotion recognition based on deep learning method. Pan [14] proposed the emotion recognition method of comment text, and used BERT to explore the emotions

in e-learning environment from different dimensions. In addition, some researchers also use explicit emotional cues to achieve emotion recognition, such as statements and words. Although utterance-based emotion recognition methods are considered to be as effective as facial expression recognition, this method has only been widely used in the online learning community, and the application in the smart classroom has been less explored.

In conclusion, the existing teacher emotion recognition work mostly use the unimodal method (such as facial expression, voice, text content, etc.) to recognize emotion in teaching scenes. Researchers seldom consider the benefits of different emotion representations on emotion recognition from the perspective of multimodal fusion. IoT-based sensors, such as cameras and microphones, embedded in smart classrooms can effectively record teaching data in real time. Based on these advantages, the smart classroom can provide a richer form of data for emotion recognition, creating a favorable environment for multimodal emotion recognition methods. The teacher utterance contains not only the implied intonation, pitch and other emotion-related features, but also the explicit textual emotional features. This inspired us to explore ways to take full advantage of the information brought by IoT devices in the smart classroom and apply a multimodal approach to emotion recognition to the smart classroom.

Multimodal speech emotion recognition

Speech contains two modalities, audio and text, which imply the emotional state of the speaker. How to obtain the implicit emotional state of the speaker from the speech has always been a research problem worth discussing. Anusha [15] utilized the syntactic and semantic features of the text to extract the implicit sentiment states. Previous studies have designed and implemented emotion recognition methods from different aspects of utterances. However, the performance of these methods is limited by the lack of information from unimodal data. Multimodal emotion recognition makes full use of the unique information from different modalities which can provide supplementary information for emotion recognition. Therefore, multimodal emotion recognition method has stronger representation ability than the unimodal emotion recognition method. The common work of multimodal emotion recognition includes: feature selection, extraction, and classification. Middya [16] et al. fused eye movement data, facial expression, speech signals and other features as the input features for emotion recognition, and constructed a lightweight multimodal fusion method. Guo [17] et al. used 3D-CNN and 1D-CNN to extract the facial features and the Mel Frequency Cepstral Coefficient (MFCC) features of audio, respectively. Guo fused these two features by Transformer module to predict the emotional results. Shou [18] constructed a multimodal relation graph based on speaker relation and dependency syntactic relation, and used the graph convolution neural network for emotion recognition. Guo [19] proposed a framework for implicit alignment of acoustic and textual features, enabling models to add multimodal information when learning emotional features. It can be seen that most of related work in emotion recognition focused on feature extraction and classification, and paid less attention to the impact of feature selection on emotion recognition.

To sum up, the existing works tend to ignore the importance of feature selection when using multimodal methods, and pay less attention to the prosodic features of the speaker. In addition, the existing work will less consider the contribution of different frames to teacher emotion. This inspired us to reflect the differences between different emotions through multiple prosodic features in the feature selection stage.

Our approach

In this section, we introduce the method we proposed. Firstly, we found and summarized the prosodic characteristics of teacher's speech emotion in smart classrooms. Secondly, a ProsodyBERT neural network that can fuse prosodic features and textual features is proposed based on the characteristics we found. The ProsodyBERT adopts a parallel coding framework, a prosody encoder and a text encoder are implemented to deeply encode the prosodic features and textual features respectively. A decoder is used to fuse the prosodic features and textual features and output recognition results.

Prosodic analysis of teacher emotion

Prosody determines the rhythm and emotional state of a sentence, it includes the change of pitch, duration and speed of words and sentences [20]. By observing lots of teaching videos, we found the characteristics of emotional prosody in teacher utterance:

(1) **Teacher emotion is deductive**. The emotions in teacher utterance are different from those in daily conversation. Compared with the emotions in daily conversation, teachers in the classroom are more "deductive" and they express their emotions more strongly

Fig. 1. Teachers show their emotion through facial expressions.

because they need to drive the learning atmosphere of students through emotions. And teachers usually pretend to be happy to arouse students' positive learning desire. The example listed in Fig. 1 shows the teachers' facial expressions when expressing specific emotions in smart classroom.

In Fig. 1, we give an example of the classroom recordings from four teachers. (a) shows that the teachers are teaching the knowledge in the textbook to the students, and the teachers are smiling but in a calm tone. (b) shows that the teacher invites the students to read the book together, smiling and in a calm tone. (c)(d) show the teacher is ready to ask questions to the students. (e) shows a teacher is asking a question to a student, hoping that the student will pay attention to his questioning instructions and pretend to look solemn. (f) shows a teacher asking questions to his students with a smile on his face.

As can be seen from (a) (b) (c) (d) (f), to mobilize students' positive emotions, teachers tend to express their emotions through smiling. In this situation, if the visual-based teacher emotion recognition method is utilized, (a) (b) (c) (d) are easy to be recognized as *Happy*. As for (e), it will be recognized as *Angry* or *Confuse*. As for (f), there is not only a teacher but also a student in the picture, which will be difficult for the visual-based method to determine the subject that needs to be recognized.

(2) **There are differences in the prosody of teacher utterance in different emotions**. To further describe this problem, we extracted and analyzed Mel-spectrogram characteristics from different teachers' utterances with different emotions. We normalized these features so that they can reflect the acoustic differences in teachers' utterances, as shown in Fig. 2.

When teachers express *Happy* emotions, they can intensify such emotions through tone changes and also through the content of utterance. As shown in Fig. 2, in (a) and (b), the *Happy* can be divided into two types. One is that the teacher's overall tone is light and relaxed. The text content in (a) is *掌声有多响, 灯就会有多亮.*(The lights turn on as loud as the applause goes.). As shown in (a), it can be seen that the energy distributes average. The text content in (b) is *哈哈😉有没有喜欢的请举手.*(Hah-hah! Please raise your hand if you like it.). It can be seen from (b) that the energy in teacher utterance is stronger than that in (a). Compared with (a), (b) highlights the teacher's pleasure.

In (c) and (d), the *Neutral* has an even energy distribution, and at the same time, the rhythm is flat and constant. In addition, there is no emotion cues in the literal expression. The text content in (c) is *如果消息传到楚王那里, 他就无法报仇了.*(If this message reaches the king of Chu, the revenge will be unavenged.). And the text content in (d) is *语文最重要的东西是人与人之间的心灵的交流.*(The most important thing in language is the communication among people's hearts.).

There are obvious differences between teachers' *Surprise* and *Happy*. Teachers will express Surprise through some expressions such as "Wow!", "Ah!", and other interjections to reinforce that feeling. In (e) and (f), the *Surprise* usually has a higher pitch than other emotions, and it can be seen that the average duration of surprise was the shortest. The text content in (e) is *真正的教育给你的绝不仅是*

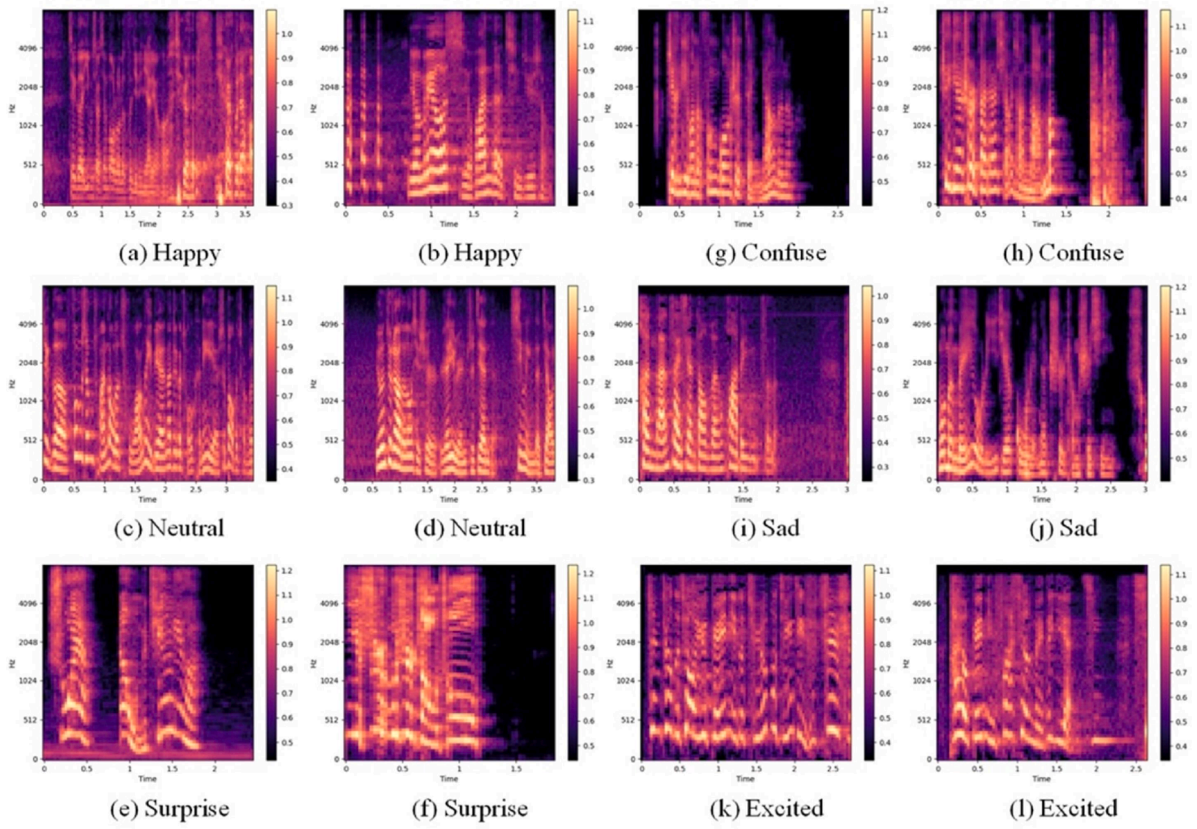


Fig. 2. Examples of audio waveforms for different emotions.

知识ঙ (Real education gives you more than just knowledge!). The text content in (f) is教育要给你发现美的眼睛.(Education gives your eyes to see beauty.).

The *Confuse* usually occurs when a teacher asks a student a question or wishes to interact with a student. Therefore, teachers have significant features in the text content when expressing *Confuse* emotions, which usually appear in the form of interrogative sentences. In most cases, the expression of *Confuse* emotion has no significant feature in acoustics (as shown in (g)). But some teachers will reinforce this emotion by re-reading (as shown in (h)), there is usually a stress in the end of utterance to achieve the purpose of emphasizing the question. The text content in (g) is 这个地方的风光是怎么样的ঌ (How is the scenery like in this place?). The text content in (h) is这件事大家想一想, 难吗ঌ(Think about it, is it difficult?).

In (i) and (j), the *Sad* has a lower average energy, and there is no obvious intonation change. Based on our observation of the data in this emotion category, we found that it may be difficult to express *Sad* emotion only in words, while the characteristics of this emotion are significant in hearing. The text content in (i) is 现当今很难再见到古文化的影子了.(It is hard to see the existence of ancient cultures nowadays.). The text content in (j) is我们难以理解古人的赤诚之心.(We can not understand the sincere heart of the ancients.).

Teachers usually express *Excited* emotions through high volume and high pitch, and at the same time, may not have significant features in the text content. The difference between *Excited* and *Surprise* is that *Surprise* may only be significant in a certain period of time in the teacher's speech, while *Excited* will always be significant. In (k) and (l), the *Excited* usually has a higher pitch, a sharper sound, multiple accents. Compared with other emotions, *Excited* has the most significant characteristics. The text content in (k) is 说呀, 让我们听听啊ঙ(Come on! Let's hear it!). The text content in (l) is 你能不能明白ঌঙ(Can you understand?!).

By observing the text contents of teachers' utterance, it can be found that, except for the *Confuse* ((g) (h)), which are obviously reflected in the text content, other categories of teachers' emotions have no obvious characteristics in the text content. Therefore, it is certain that when teachers do not express emotions through specific modal words, such as "Hah-hah" in *Happy* (b), it is difficult to correctly recognize teachers' speech emotions only through the content of utterance.

Data preprocess

Firstly, for the teaching videos collected from the smart classroom, the teachers' speeches are separated from the teaching videos. Subsequently, the speeches were segmented and removed the silent part with voice activity detection (VAD) based on short-time energy and zero crossing rate [21]. VAD segments speech by detecting silent segments in speech. The purpose of using VAD is to eliminate silent segments that do not contain any emotional information, and will cause greater information redundancy [22]. In addition, VAD is not completely accurate, so it requires a longest and shortest hyperparameter to mark the appropriate segment range, and the range which is too long or too short can also cause VAD errors. To obtain more exact speech segments, we observed lots of fully expressed and clearly articulated teacher utterances and found that the average length of teachers' utterance was about 1.5–10 s. so we limited the minimum and maximum length of each segmented fragment to 1s-12 s. The speech segment can be represented as $S = \{S_1, S_2, \ldots, S_n\}$.

Teachers' speeches are transcribed into text with Baidu's auto speech transcription API, the transcribed text can be represented as $S_t = \{S_1, S_2, \ldots, S_n\}$. Since BERT for classification tasks requires a $[CLS]$ token to represent the start of the input, text sent to the BERT model can be represented as $S_t = \{S_{[CLS]}, S_1, S_2, \ldots, S_n\}$.

Manually crafted features such as Mel-spectrogram, Mel-scale Frequency Cepstral Coefficients (MFCC), and Filter-bank are widely used as frequency domain features of speech in automatic speech recognition (ASR) and speaker diarization (SD) [23,24]. These features convert the speech signal from the time-domain to the frequency-domain, which is consistent with the characteristics of human hearing [25]. Compared with Filter-bank and MFCC, Mel-spectrogram features have strong robustness and speech discrimination, thus speech recognition related studies tend to choose Mel-spectrogram as speech feature [26]. To obtain the Mel-spectrogram features of teacher utterances, the speeches are uniformly processed to 16 kHz at first. The audio is pre-emphasis to compensate for the energy loss during sound propagation, and the audio is divided into 25 ms short frames, adding a Hamming window with 10 ms shift to the speech signal. Subsequently, the audio is converted from time-domain to frequency-domain by Fourier transform, and the spectrum is passed through a set of Mel-scale filter banks and log-Mel filter is obtained by logarithmic operation. Finally, the Mel-spectrogram features with 50 dimensions are obtained.

Only Mel-spectrogram is difficult to fully reflect the prosodic features of utterance [27]. Research has shown that prosodic features can more fully reflect the emotion of utterance [28]. Using prosodic features from different domains can provide more effective information for TSER. Therefore, the prosodic related features of different domains are added, such as time-domain, frequency-domain, energy-domain, and perception-domain (Table 1). The prosodic features are extracted by the method given in *librosa* library of Python 3. It is worth noting that no more tricks were used when merging these prosodic features with Mel-spectrogram. We placed prosodic

Table 1
Description of the prosodic feature.

Domain	Feature and description
Time	**Attack time**: Duration of sound energy in the rising phase.
	Zero cross rate: The number of times the signal passes through the X-axis, it reflects the frequency characteristics of audio signals.
Frequency	**Spectral centroid**: The brightness of the speech. The lower the spectral centroid, the darker and deeper the speech.
Energy	**Root mean square**: The mean energy of speech over a certain time period.
Perception	**Sharpness**: The sharpness of the sound.

features after Mel-spectrogram in the feature dimension by a simple concatenation operation, and limited the dimension of these prosodic features to 30, so the dimension size after concatenation is 80. Although the teacher emotion is also related to the length of the speech, the length information of the speech is included in the audio itself, so no additional processing was performed.

Multimodal alignment is regarded as an important work in multimodal methods, which can effectively improve the utilization of multimodal information [29]. To make the audio and text more compatible, we refer to the work in [30] and add a zero vector before the audio so that the zero vector acts as a placeholder to fill the [CLS] token in Bidirectional Encoder Representation from Transformers (BERT) [31]. Then, the audio input is represented as $S_a = \{S_{[CLS]}, S_1, S_2, \ldots, S_n\}$, where $S_{[CLS]}$ denotes the zero vector before audio input.

ProsodyBERT

The processed audio and text from Section 3.2 are fed into ProsodyBERT. ProsodyBERT adopts an encoder-decoder network structure with a text encoder, an audio encoder, and a decoder.

The structure of ProsodyBERT is shown in Fig. 3. Firstly, the text encoder and prosody encoder receive the manually crafted features and send the features to the decoder after deep encoding. Then, the decoder decodes the fused features. Finally, the decoder outputs teacher speech emotion category.

Text encoder

BERT is chosen as the encoder for the text content in teacher utterance (Corresponding to Fig. 3(d)). BERT is a large pre-trained model composed of multiple Transformer [32] architectures and trained on massive of prior knowledge. BERT has excellent natural language representation ability; it can provide language representation support for downstream tasks. The version of BERT pre-trained model is Chinese $\text{BERT}_{\text{BASE}}$, which contains 12 transformer blocks and 110 M parameters. In the text encoder, BERT is used to obtain the high-dimensional feature x_{text} of the text, which has 768 dimensions.

Prosody encoder

The prosody encoder includes an 1D-convolution block, attentive pooling, and fully connected block (Corresponding to (a) (b) (c) in the Fig. 3, respectively).

The 1D-convolution block contains two 1D-convolution layers. In the 1D-convolution block, to more fully extract teacher emotional information at different scales, we adopted convolution kernels at different scales to extract teacher utterance features. The first 1D-convolution layer receives the input feature S_a and captures the coarse-grained features x_a, where the input channel is 80, the output channel is 512, and the kernel size is 5.

$$x_a = \text{Norm}(\text{RELU}(\text{Conv1D}(S_a))) \tag{1}$$

where Norm denotes 1D Batch-normalization. Subsequently, x_a is transmitted to the second layer with both input and output channels are 512 and kernel size is 1.

The attention pooling layer receives the output x_a and calculate the weighted mean and weighted standard deviation to selects the

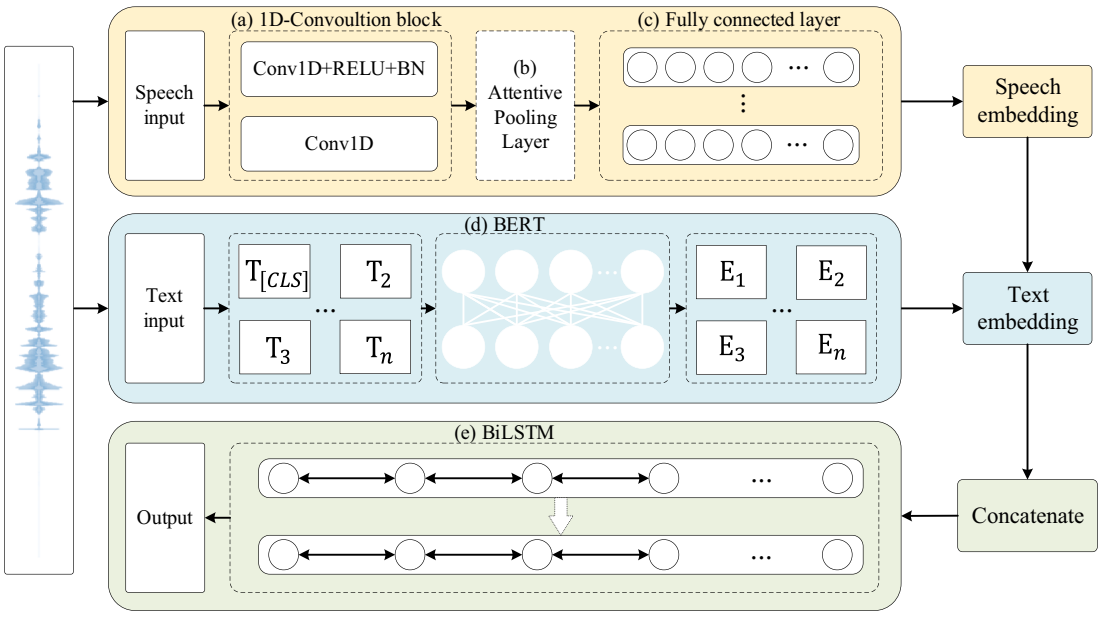


Fig. 3. The structure of ProsodyBERT.

key features. Firstly, the key features were calculated as scalar score s_t

$$s_t = v^t f(Wh_t + b) + k \tag{2}$$

where $f(\cdot)$ is a non-linear activation function, and we used *tanh* here. $h_t (t = 1, ..., T)$ are the activations of the last frame layer at time step t. We used the Softmax function to normalize s_t and obtain the attention score α_t, and α_t is the weight of pooling layer to calculate the weighted average vector μ

$$\mu = \sum_{t}^{T} \alpha_t h_t \tag{3}$$

Then, we adopted the method proposed by Okabe [33] to calculate the weighted standard deviation σ to make the attention pooling layer pay more attention to the valuable speech frames, and the calculation is

$$\sigma = \sqrt{\sum_{t}^{T} \alpha_t h_t \odot h_t - \mu \odot \mu} \tag{4}$$

where $\odot$ represents the Hadamard product. The final output of attentive pooling is given by concatenating μ and σ, both of which have 512 dimensions. Therefore, the final output of attention pooling is the sum of the μ and σ, which is 1024.

The fully connected block include two fully connected layers is implemented to extract the utterance-features from the output of attentive pooling, the first fully connected layer is used to reduce the feature dimension to 512. Subsequently, the second fully connected layer receives the output and keeps the features at their original dimension 512. The output of prosody encoder can be represented as x_{audio}.

Decoder

A decoder is deployed to receive and decode the output from the prosody encoder and text encoder, as shown in Fig. 2(e). In the decoder, we integrated the text and prosody of teacher utterance. Since the features were previously selected by attention in both the prosody encoder and the text encoder, we only adopted the concatenation operation when fusing the text and audio modes in the decoder. Before the audio and text embeddings sent to the decoder block, the audio and text features are concatenated in the feature dimension, the fused feature is represented as H_t.

$$H_t = Concatenate(x_{text}, x_{audio}) \tag{5}$$

The dimension of H_t is 1280. H_t are regarded as a sequence in decoder, and as recurrent neural network, a GRU has an extremely strong ability to extract the time dependence in the sequence [34]. Therefore, two layers of bidirectional Gate Recurrent Unit (GRU) are stacked to capture the feature dependencies of the time series in the fused features of teacher utterance. The dimension of hidden layer is 256. To avoid overfitting, 0.3 dropout is set between the hidden layers. $c_i^s \in \mathbb{R}^{d_u}$ represents each fused feature, The output $H_i^s \in \mathbb{R}^{2d_u}$, which characterized by GRU can be expressed as:

$$H_i^s, h_i^s = \overleftrightarrow{GRU}\left(c_i, h_{i-1}^s\right) \tag{6}$$

where $h_i^s \in \mathbb{R}^{d_u}$ denotes the i-th hidden layer state of GRU. Then the final teacher emotion category is obtained.

Experimental results and discussion

In this section, we describe the experiment on ProsodyBERT. In subsection 4.1, we introduce the datasets we adopted and the experimental environment. Subsection 4.2 explains the measurement criteria of the experimental results on different datasets, subsection 4.3 introduces the specific setting of model parameters and the experimental methods adopted. In subsection 4.4, the comparison with other best methods on IEMOCAP and MELD datasets proves the effectiveness and superiority of our proposed method, and experiments on the MTED prove the feasibility of the method in smart classroom scenes. Then, the contribution of different modalities and input features to the recognition of teacher emotion was explored through ablation experiments. We compared our method with unimodal methods based on text and facial expression in subsection 4.5. In subsection 4.6, we compared the performance of our method with other unimodal methods in videos recorded by smart classrooms. In subsection 4.6, we discussed the results obtained from the experiments.

Experimental datasets and environment

We used the Interactive Emotional Motion Capture (IEMOCAP) [35] and Multimodal Emotion-Lines Dataset (MELD) [36] as experimental datasets to evaluate the performance of our proposed method. To explore the availability of our proposed method in TSER task, a dataset named Multimodal Teacher Emotion Dataset (MTED) is constructed in this work.

IEMOCAP

The Interactive Emotional Motion Capture (IEMOCAP) dataset was developed for SER. The IEMOCAP dataset including scripted and impromptu emotional speeches, divided into five sessions. The IEMOCAP dataset contains three modalities of data: video, audio and text, including 10 emotions in total. The 4 types commonly used for experiments are *angry, happy, neutral*, and *sad*. The data distribution is shown in Table 2.

MELD

The data in Multimodal Emotion-Lines dataset (MELD) has three modalities: video, audio and text. The MELD contains more than 1400 dialogues and 13,000 quotes from the *Friends* TV series. Several speakers took part in the dialogue. The distribution of MELD has shown in Table 3.

MTED

Previous studies paid little attention to the impact of utterance-level information brought by smart classrooms on TSER tasks, resulting in a lack of relevant datasets. In order to evaluate the effect of our proposed method in smart classroom, we collected 235 high-quality classroom videos recorded by the smart classroom and processed the video data to generate the MTED dataset. MTED includes three subjects, biology, information technology and Chinese. Each class has at least one teacher, so there are a total of 235 teachers' voices in MTED. Two trained students of educational technology were assigned to manually label the dataset (text and audio). The consistency of the data annotation was determined by calculating the Kappa value, which yielded a result of 0.77, signifying a high degree of consistency. MTED divided teacher emotions into 6 categories: *Excited, Happy, Neutral, Confused, Sad* and *Surprise*. The data distribution of MTED is shown in Table 4.

Experimental environment

The hardware environment of this experiment is Intel i7–11700k @3.6 GHz and NVIDIA 3080Ti; the software environment is Pytorch 1.10.1 deep learning framework built under the 64-bit Ubuntu 22.04 operating system.

Evaluation metrics

Unweighted accuracy (*UA*) and weighted average F1-score (*W-Avg F1*) are widely used to evaluate the effect of models on multimodal datasets such as IEMOCAP, MELD. To evaluate the effects of ProsodyBERT on the IEMOCAP, MELD, and MTED datasets, and to compare our methods with methods proposed in previous studies, we adopted the evaluation metrics that have been used in previous studies. 4-class unweighted accuracy (UA_4) and weighted average F1-score are used as the evaluation metrics for IEMOCAP, and 6-class unweighted accuracy (UA_6) and *W-Avg F1* are used for MTED. The *UA* is calculated as:

$$UA = \frac{\sum_{i=1}^{N} t_{ii}}{\sum_{i=1}^{N}\sum_{j=1}^{N} t_{ij}} \tag{7}$$

where N represents the number of categories, t_{ij} denotes the number labeled as category i but predicted as category j. The weighted Average F1 score is calculated as:

$$\mathrm{W-Avg\ F1} = \frac{1}{N}\sum_{i=1}^{N}\frac{2Precision \cdot Recall}{Precision + Recall_i} \tag{8}$$

We used 6-class unweighted accuracy (UA_6) and *W-Avg F1* as the evaluation metrics for the evaluation of MELD. 6-class unweighted accuracy (UA_6) and *W-Avg F1* are used for the evaluation of MTED. The data for different modalities (text, audio, video) are represented as: *T, A, V*.

Implementation details

We fine-tuned the ProsodyBERT on the validation set of IEMOCAP and MELD. We continuously adjust the parameters in the validation set of IEMOCAP, and also refer to previous experiences [37] to ensure that the model can perform well in the SER tasks.

Table 2
The distribution of emotion in IMEOCAP.

Emotion	Total
Anger	1103
Happy	1689
Neutral	1708
Sad	1084
Total	5584

Table 3
The distribution of MELD.

Emotion	**Total**
Anger	1256
Disgust	293
Sadness	794
Happy	1905
Neutral	5178
Surprise	1355
Fear	308
Total	11,089

Table 4
The distribution of emotion in MTED.

Emotion	**Total**
Happy	1041
Surprise	565
Neutral	1863
Sad	411
Excited	334
Confuse	1749
Total	5963

Through adjusting the hyperparameters, the batch size is 64, learning rate is 5e-6, and the Adam optimization strategy is adopted, the total training epoch was 30. It should be noted that since IEMOCAP and MELD are English datasets, the pre-trained $BERT_{BASE}$ model is used as the text encoder instead of using the Chinese $BERT_{BASE}$ uncased. All the experiments reported in this work used the same hyperparameters.

10-fold cross-validation method was used for experiments on IEMOCAP, MELD and self-built MTED dataset. We divided the data in each dataset into 10 folds, one for testing and the remaining nine for testing. In each fold on IEMOCAP, MELD, and MTED, the training time averaged from 18 to 20 min.

Experimental results

Experiment on IEMOCAP and MELD

IMEOCAP and MELD public datasets are used to evaluate the effect of our proposed method. Several state-of-the-art methods on the IEMOCAP are selected as the benchmark model to compare with our method:

DBT [38]. RoBERTa is used to encode text features, wav2vec 2.0 is used to extract audio features, and a weighted label smoothing strategy is proposed, which effectively improves the accuracy of emotion recognition.

CHFusion [39]. A novel fusion strategy was implemented to fuse modalities in pairs, which proved to be effective in multimodal

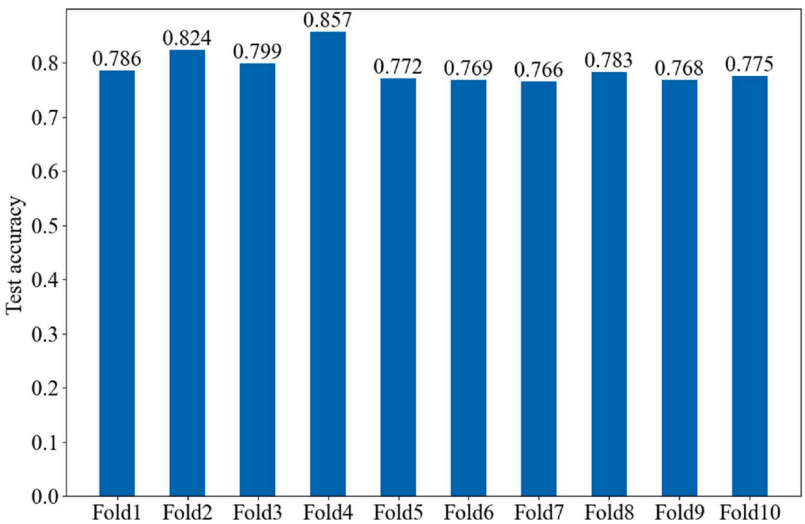


Fig. 4. The test accuracy of ProsodyBERT on IEMOCAP.

emotion recognition task.

Makiuchi [40]. Wav2vec 2.0 was used to extract audio features and speaker information was added to the audio. Based on the BERT model, the textual features of the utterance are obtained and combined together to obtain the sentiment classification.

SAWC [41]. A method is proposed to apply attention mechanisms and confidence measures to SER tasks. It makes SER reduce the dependence on text features.

On the MELD, serval benchmark models are selected to compare the performance of our proposed method: bcLSTM [42]. Based on LSTM, the model is enabled to capture contextual information from the surrounding environment in the same video.

DialogueGCN [43]. The conversational context is simulated using the dependencies between the speaker and the interlocutor.

DialogueCRN [44]. Reasoning modules is proposed to obtain and integrate emotional information to perceive emotional context.

MM-DFN [45]. A multimodal dynamic fusion method is proposed that can fully understand dialogue context.

Fig. 4 and Fig. 5 depicts the results of each fold when testing ProsodyBERT on the IEMOCAP and MELD datasets using 10-fold cross-validation. By observing the data, it can be found that the performance of ProsodyBERT on IEMOCAP is significantly better than that on MELD.

As shown in Table 5, our proposed ProsodyBERT surpasses current state-of-the-art methods and achieves the best performance on IEMOCAP. The UA_4 and the *W-Avg F1* are 78.6 % and 77.8 % respectively. The proposed ProsodyBERT achieves better performance on MELD compared to several state-of the-art methods. The UA_6 and the *W-Avg F1* are 66.2 % and 64.7 % respectively.

It can be seen that our proposed method has better performance on IEMOCAP, which is consistent with the results obtained by some other related studies. Firstly, this may be because IEMOCAP with four emotional categories and MELD with six emotional categories were used, with fewer tasks for IEMOCAP. Secondly, IEMOCAP is more suitable for daily conversation, while the data in MELD are all from the *Friends*, so it is more inclined to stage scenes.

Experiment on MTED

According to the characteristics of teachers' emotional utterance obtained in Chapter 3.1, it can be found that the expression of emotions in teacher utterance is different from daily utterance. As a result, the existing unimodal emotion recognition method may not be suitable for the actual teaching situation. We implemented experiments on the MTED to examine the effect of our proposed method in the smart classroom.

Fig. 6 describes the results of each fold when testing ProsodyBERT on MTED. As can be seen from Fig. 6, ProsodyBERT achieved 82.1 % Acc_6 and 81.7 % *W-Avg F1* on the MTED dataset. The result also shows that our ProsodyBERT can be effectively applied to the TSER tasks in smart classroom scenes.

Ablation experiments

In order to compare the contribution of multimodal and multiple prosodic features to TSER, we designed and performed the ablation experiment. In the ablation experiment, the modules such as prosody encoder, text encoder and multiple prosodic features were eliminated successively. By comparing the statistical table and confusion matrix of the results, we explored the influence of different modules on the performance of ProsodyBERT on TSER task. The procedure and description of the ablation experiment are as follows:

(a) **ProsodyBERT with multiple prosodic features:** Originally designed ProsodyBERT.

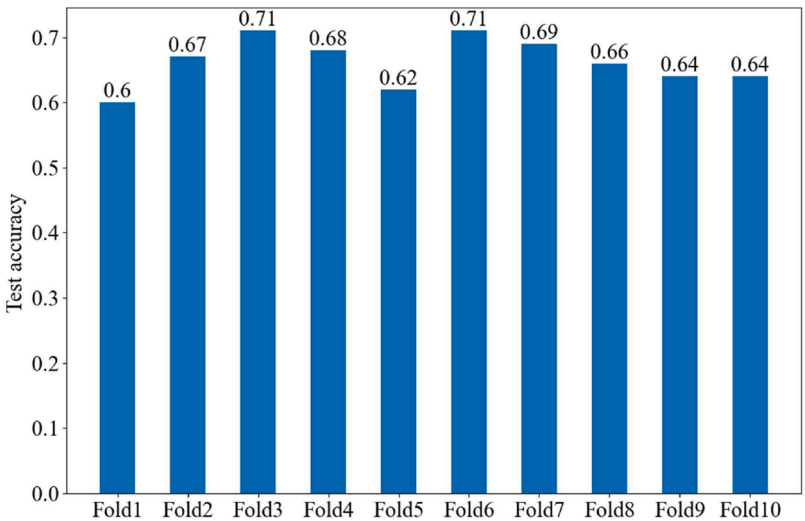


Fig. 5. The test accuracy of ProsodyBERT on MELD.

Table 5
Comparison of experimental results on IEMOCAP and MELD.

Method	Modality	Dataset	UA_4. (%) / UA_6. (%)	*W-Avg F1*. (%)
DBT	*T & A*	IEMOCAP	74.0	/
CHFusion	*T & A & V*	IEMOCAP	76.5	76.8
Makiuchi	*T & A*	IEMOCAP	73.2	/
SAWC	*T & A*	IEMOCAP	76.8	76.9
ProsodyBERT (Ours)	***T & A***	IEMOCAP	**78.6**	**78.8**
bc-LSTM	*T & V*	MELD	57.5	56.4
DialogueGCN	*T*	MELD	59.5	58.1
DialogueCRN	*T*	MELD	60.7	58.4
MM-DFN	*T & A*	MELD	62.5	59.5
ProsodyBERT (Ours)	***T & A***	MELD	**66.2**	**64.7**

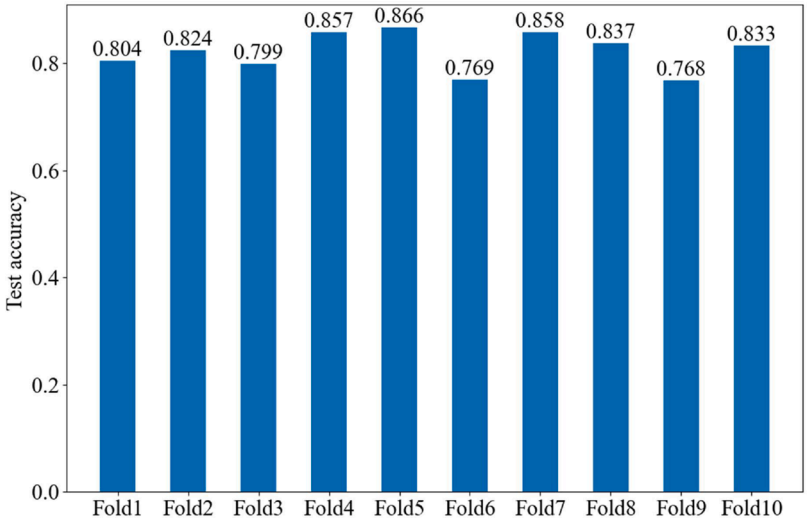


Fig. 6. The variation trend of training loss and test accuracy.

(b) **ProsodyBERT with Mel-spectrogram only:** All the components of ProsodyBERT are retained, but only MFCC are selected for the features of the audio modality.
(c) **Replacing attention pooling layer:** Replacing the attention pooling layer in the prosody encoder with a max pooling layer, which is the most commonly used pooling strategy in CNN.
(d) **Replacing decoder:** Replacing the GRU with a multilayer perceptron.
(e) **Only text encoder:** The text encoder was retained and all components were removed.
(f) **Only prosody encoder:** The prosody encoder was retained and all components were removed.

The confusion matrix (Fig. 7) reflects the performance of ProsodyBERT on TSER task after continuously removing different modules.

Evaluation in practical smart classroom

We obtained 10 teaching videos recorded by smart classrooms outside the MTED dataset to compare the practical effect of our proposed method in smart classroom environments. After VAD segmentation, there are 1732 teachers' voice data. We compared our proposed method with text-based methods. We compared the bidirectional long term memory network (Bi-LSTM) and the pre-trained large model ERNIE [46] proposed by Baidu with our method. Among them, ERNIE is optimized specifically for Chinese context based on BERT, and has achieved excellent results in Chinese-related downstream tasks. All the models used for testing were trained on the MTED dataset, and the indicators used for evaluation are the average UA_6, *W-Avg F1* and the average running time. The comparison results are shown in the Table 7.

It can be seen from Table 7 that ProsodyBERT requires the highest time cost in the practical smart classroom environment, but at the same time, ProsodyBERT performs significantly better than text-based emotion recognition methods. After analysis, we speculated that the reason for the longest running time of ProsodyBERT may be due to the additional processing of audio signals, which increases the computational difficulty.

In practical application, public speech transcription tools often have transcription errors due to network problems. To evaluate the

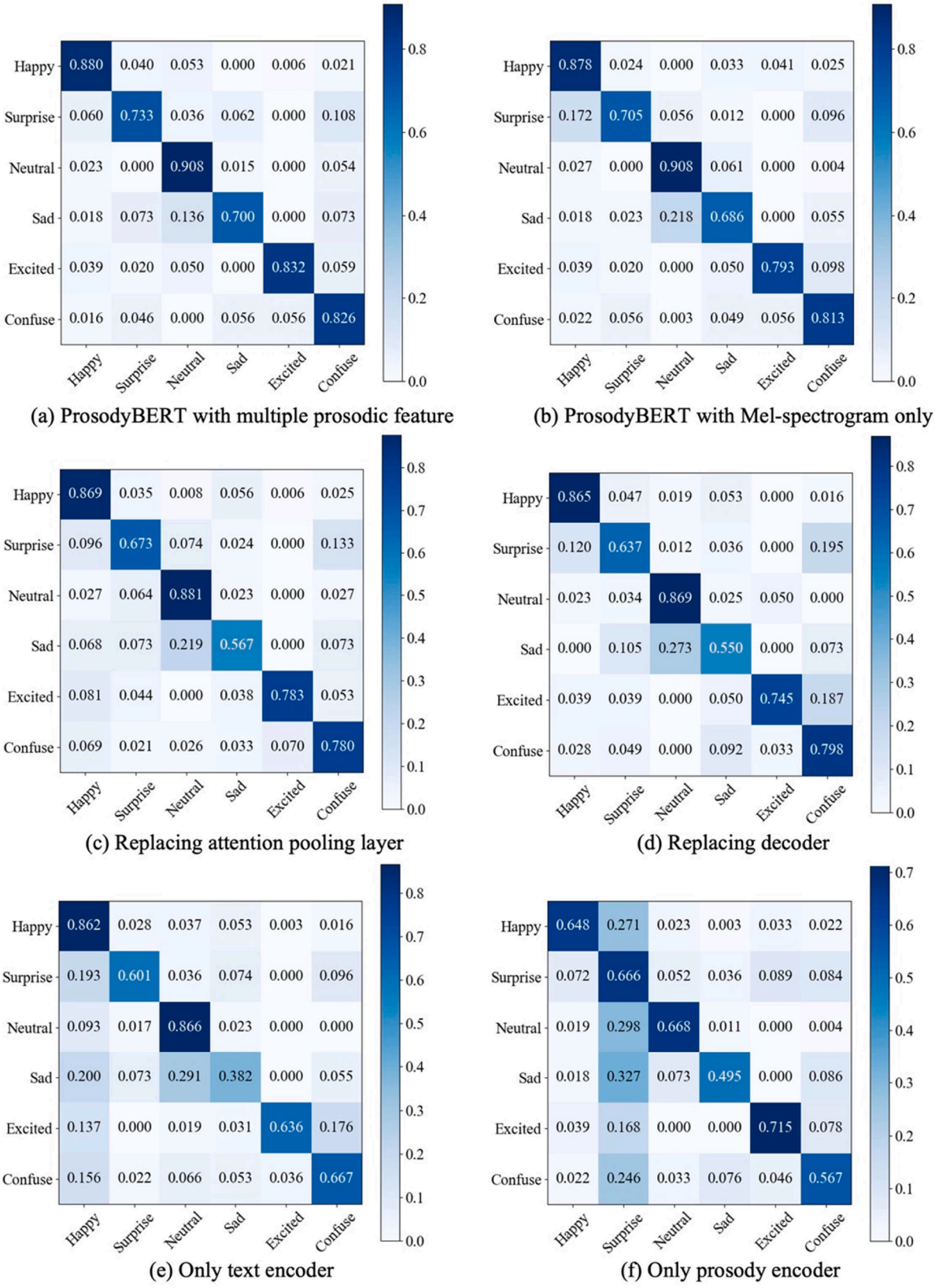


Fig. 7. Comparison of normalized confusion matrixes for ProsodyBERT.

robustness of our proposed method in this case, we randomly masked the transcribed text used for testing at 10 %, 20 %, and 30 % by replacing it with an [*Error*] token.

To compare the loss of model performance at different damage levels, we introduced tests without any masking for reference. As shown in Table 8, the UA_6 of Bi-LSTM, ERNIE, and ProsodyBERT is decreased by 5.3 %, 4.3 %, and 3.2 % when 10 % is masked, and the gap is small. The UA_6 decreased by 13.2 %, 12.6 %, 7.6 % respectively when 20 % is masked, and by 21.8 %, 19.2 %, 14.3 % respectively when 30 % is masked. It can be seen that compared with text-based emotion recognition methods, ProsodyBERT has stronger robustness in the face of the loss of transcribed text due to network problems. This result also illustrates the importance of complementary audio and text modal information in TSER tasks.

Discussion

The ablation experiments better explained the contribution of different modules to teacher emotion recognition in our proposed method. By observing the results in Table 6, it can be seen that the addition of multi-prosodic features improves UA_6 and *W-Avg F1* by 1.2 % and 0.9 %, respectively. By comparing (a) and (b) in Fig. 7, it can be seen that multi-prosodic features play a certain role in improving the performance of teacher emotion recognition. Among them, the category most affected by multi-prosodic features is *Sad*, with an improvement of 6.4 %.From Table 6 (c), when attention pooling is replaced by max pooling, UA_6 and *W-Avg F1* are reduced by 1.2 %and 1.1 % respectively, which proves to some extent that the attention pooling layer makes the model pay more attention to the most valuable audio frames in prosodic features and reduces the feature redundancy. By comparing (b) and (c) in Fig. 7, it can be found that attention pooling structure is of great help to identify different teacher emotions, and the *Confuse* category is the biggest influence of attention pooling. The *Confused* was the most affected by attention pooling, and attention pooling brought 3.3 % improvements to the recognition of this category. As can be seen from Table 6, replacing the proposed structure in the decoder module with multilayer perceptron reduces UA_6 and *W-Avg F1* by 1.1 % and 1.7 %, respectively. Comparing (c) and (d) in Fig. 7, it can be found that replacing the GRU structure in decoder with a fully connected layer affects the recognition effect of all categories, among which *Surprise* and *Anger* have the largest reduction of 3.5 % and 3.8 %, respectively. This result indicates that the GRU structure we are using can effectively capture the sequence dependencies of prosodic and textual features. From Table 6, it can be found that when only the text encoder is retained, UA_6 and *W-Avg F1* decrease by 6.6 % and 7.3 %. This experiment simulates the absence of audio modality and reflects the influence of text modality on teacher emotion recognition. By comparing (d) and (e) in Fig. 7, it can be found that the recognition effect of *Sad* is most affected when only the text encoder is retained, which is reduced by 30.9 %, and the *Sad* is more inclined to be recognized as *Neutral*. It confirms the prosodic characteristics of *Sad* we summarized in Section 3.1, that is, *Sad* emotion is difficult to recognize only through text features. It can be found from (d) and (f) in Table 6, when only the prosody encoder is retained, UA_6 and *W-Avg F1* decrease by 18.4 % and 19.2 %. It simulates to a certain extent that the absence of text modality, and indicates that the effect of teacher emotion recognition is greatly reduced when only audio modality is used to recognize the teacher emotion. By comparing (d) and (f) in Fig. 7, the performance of the text encoder only being slightly better than that of the prosody encoder only, indicating that text modality seems to contain more emotion information. However, prosody encoder performs better than only text encoder in recognizing *Anger, Surprise* and *Sad*.

Conclusion

Aiming at TSER in smart classroom, based on the analysis of the differences in prosodic characteristics of teacher emotions, we proposed a method to enhance the model's ability to perceive emotions by adding multiple prosodic features. Benefiting from the rich data forms provided by the IoT-based smart classroom, this paper designed and proposed the ProsodyBERT neural network, which not only considered the prosodic features and text features in speech, but also selected the key prosodic features through attention pooling. The experimental results on IEMOCAP and MELD show that our method outperforms the existing methods and prove the advancement of our proposed method. The experimental results on MTED dataserecorded by the smart classroom show that our proposed method can more effectively recognize teachers' speech emotion in smart classroom scenes. The main contributions are:

(1) This paper makes full use of the information advantage in smart classroom. We found and summarized the emotional prosodic features in teachers' speech, and proposed a teacher emotion recognition method based on the fusing prosodic features and textual features. This paper aims to provide a new and reliable emotion recognition tool for teachers' emotion analysis, and analyzes and captures the prosodic features of teachers' utterance in the wisdom classroom, so as to realize teachers' emotion recognition. But it does not mean that these prosodic features are unique to teachers, and may also be applied to relevant tasks where the speaker is in a specific context.

(2) In view of the problem that the teacher facial expression may not reflect the true teacher emotion, the existing methods mainly use unimodal method to recognize the teacher's emotion. This paper proposed a ProsodyBERT that can integrate multimodal and multiple prosodic features. ProsodyBERT makes full use of information in teacher utterance, providing a feasible way for teacher emotion recognition in smart classroom.

There are still some shortcomings in this study: (1) Although the original intention of this study is to provide a new recognition tool for teacher emotion analysis in smart classroom. However, in some countries or regions, teachers are facing increasing pressure from career development, and the observation of teachers' classroom data may aggravate teachers' job burnout stress. (2) We have proposed a method for TSER that does not rely on computer vision at all, but this does not mean that this method has obvious superiority. The smart classroom can provide not only a large amount of auditory information, but also multi-angle visual information, and combining auditory and visual information in actual TSER tasks seems to be a more effective strategy.

Table 6
The results of the ablation experiment.

Step	Module	UA_6. (%)	*W-Avg F1*. (%)
(a)	ProsodyBERT with multiple prosodic features	82.1	81.7
(b)	ProsodyBERT with Mel-spectrogram only	81.5	81.3
(c)	Replacing attention pooling layer	80.3	80.2
(d)	Replacing decoder	79.1	78.5
(e)	Only text encoder	72.5	71.2
(f)	Only prosody encoder	60.7	59.3

Table 7
The comparison results of record from smart classroom.

Method	UA_6	*W-Avg F1*	Running time
Bi-LSTM	55.7 %	53.6 %	3 min **55s**
ERINE	62.6 %	61.5 %	4 min 21s
ProsodyBERT (Ours)	**71.6 %**	**70.3 %**	6 min 04s

Table 8
The test results in different masking proportions.

Proportions of masked	Bi-LSTM		ERNIE		ProsodyBERT	
	UA_6	*W-Avg F1*	UA_6	*W-Avg F1*	UA_6	*W-Avg F1*
Base-line (0 %)	55.7 %	53.6 %	62.6 %	61.5 %	**71.6 %**	**70.3 %**
10 %	50.4 %	50.1 %	58.3 %	56.2 %	**68.4 %**	**67.1 %**
20 %	42.5 %	40.6 %	50.0 %	48.8 %	**64.0 %**	**62.6 %**
30 %	33.9 %	32.1 %	43.4 %	42.7 %	**57.3 %**	**56.1 %**

As a typical case of the innovative application in IoT, smart classroom has been applied in most areas. However, the supporting software services still need to be further explored. There seems to be some research value in exploring AI to empower IoT devices in smart classrooms. In the current work, we found emotion in the teacher utterance has a strong connection with the teaching purpose. For example, before teachers want to ask students questions, they often express doubts to enhance students' interest in the problem. The teaching purpose can be obtained through the teachers' speech acts. Therefore, in future research, we expect to regard teachers' speech act as an important clue in recognizing teachers' emotion and explore the influence of behavior on emotion recognition.

CRediT authorship contribution statement

Gang Zhao: Supervision, Resources, Project administration, Funding acquisition, Conceptualization. **Yinan Zhang:** Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Formal analysis, Data curation. **Jie Chu:** Supervision, Data curation.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The authors do not have permission to share data.

Acknowledgments

This study was supported by National Natural Science Foundation of China (Grant No. 62377020) named "Research on the Intelligent Evaluation Method of Teacher Information Technology Application Ability based on Classroom Teaching Behavior Characteristics", Fundamental Research Funds for the Central Universities named "Research on Intelligent Decision-making and Service Platform for ** based on Knowledge Graph" (CCNU22JC027), and "Research on Intelligent Evaluation Technology and Application of Teachers' Teaching Ability based on Multimodal Data" (CCNU22JC011).

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.iot.2024.101069.

References

[1] E. Uzuntiryaki-Kondakci, Z.D. Kirbulut, E. Sarici, O. Oktay, Emotion regulation as a mediator of the influence of science teacher emotions on teacher efficacy beliefs, Educ. Stud. 48 (5) (2022) 583–601.
[2] L. Jie, Z. Xiaoyan, Z. Zhaohui, Speech emotion recognition of teachers in classroom teaching, in: 2020 Chinese Control and Decision Conference (CCDC), IEEE, 2020, pp. 5045–5050.
[3] S. Zhang, C. Li, Research on feature fusion speech emotion recognition technology for smart teaching, Mobile Information Systems (2022) 2022.
[4] F. Lv, X. Chen, Y. Huang, L. Duan, G. Lin, Progressive modality reinforcement for human multimodal emotion recognition from unaligned multimodal sequences, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 2554–2562.
[5] 20226472-6485 G.N. Dong, C.M. Pun, Z. Zhang, H. Anttila, K. Pyhältö, T. Soini, J. Pietarinen, Temporal relation inference network for multimodal speech emotion recognition, IEEE Transactions on Circuits and Systems for Video Technology 32 (9) (2016) 451–473, 6472-6485How does it feel to become a teacher? Emotions in teacher education. Social Psychology of Education, 19.
[6] H. Anttila, K. Pyhältö, T. Soini, J. Pietarinen, How does it feel to become a teacher? Emotions in teacher education, Social Psychology of Education 19 (2016) 451–473.
[7] A.C. Frenzel, L. Daniels, I. Burić, Teacher emotions in the classroom and their implications for students, Educ. Psychol. 56 (4) (2021) 250–264.
[8] D. Dukić, A. Sovic Krzic, Real-time facial expression recognition using deep learning with application in the active classroom environment, Electronics. (Basel) 11 (8) (2022) 1240.
[9] B. Gao, Application of Convolutional Neural Network in Emotion Recognition of Ideological and Political Teachers in Colleges and Universities, Scientific Programming, 2022, p. 2022.
[10] Z. Zhu, X. Zheng, T. Ke, G. Chai, Emotion Recognition in Learning Scenes Supported by Smart Classroom and Its Application, Traitement du Signal 40 (2) (2023).
[11] ... & S. Fakhar, J. Baber, S.U. Bazai, S. Marjan, M. Jasinski, E. Jasinska, S. Hussain, Smart classroom monitoring using novel real-time facial expression recognition system, Applied Sciences 12 (23) (2022) 12134.
[12] Y. Dai, Foreign Language Teachers' Emotion Recognition in College Oral English Classroom Teaching, Front. Psychol. (2021) 5139.
[13] F. Wang, AI-based English teaching cross-cultural fusion mechanism, Evol. Intell. (2022) 1–7.
[14] X. Pan, B. Hu, Z. Zhou, X. Feng, Are students happier the more they learn?–Research on the influence of course progress on academic emotion in online learning, Interactive Learning Environments (2022) 1–21.
[15] V. Anusha, B. Sandhya, A learning based emotion classifier with semantic text processing. Advances in Intelligent Informatics, Springer International Publishing, 2015, pp. 371–382.
[16] A.I. Middya, B. Nag, S. Roy, Deep learning based multimodal emotion recognition using model-level fusion of audio–visual modalities, Knowl. Based. Syst. 244 (2022) 108580.
[17] P. Guo, Z. Chen, Y. Li, H. Liu, Audio-visual fusion network based on conformer for multimodal emotion recognition, in: In Artificial Intelligence: Second CAAI International Conference, CICAI 2022, Beijing, China, August 27–28, 2022, pp. 315–326. Revised Selected Papers, Part II (pp.Cham: Springer Nature Switzerland.
[18] Y. Shou, T. Meng, W. Ai, S. Yang, K. Li, Conversational emotion recognition studies based on graph convolutional neural networks and a dependent syntactic analysis, Neurocomputing. 501 (2022) 629–639.
[19] L. Guo, L. Wang, J. Dang, Y. Fu, J. Liu, S. Ding, Emotion Recognition With Multimodal Transformer Fusion Framework Based on Acoustic and Lexical Information, IEEE MultiMedia 29 (2) (2022) 94–103.
[20] J. Hirschberg, Communication and prosody: functional aspects of prosody, Speech. Commun. 36 (1–2) (2002) 31–43.
[21] X. Yang, B. Tan, J. Ding, J. Zhang, J. Gong, Comparative study on voice activity detection algorithm, in: 2010 International Conference on Electrical and Control Engineering, IEEE, 2010, June, pp. 599–602.
[22] M. Faghani, H. Rezaee-Dehsorkh, N. Ravanshad, H. Aminzadeh, Ultra-low-power voice activity detection system using level-crossing sampling, Electronics. (Basel) 12 (4) (2023) 795.
[23] M.B. Akçay, K. Oğuz, Speech emotion recognition: emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers, Speech. Commun. 116 (2020) 56–76.
[24] A. Koduru, H.B. Valiveti, A.K. Budati, Feature extraction algorithms to improve the speech emotion recognition rate, Int. J. Speech. Technol. 23 (1) (2020) 45–55.
[25] K.S.R. Murty, B. Yegnanarayana, Combining evidence from residual phase and MFCC features for speaker recognition, IEEe Signal. Process. Lett. 13 (1) (2005) 52–55.
[26] H. Zhou, D. Meng, Y. Zhang, X. Peng, J. Du, K. Wang, Y. Qiao, Exploring emotion features and fusion strategies for audio-video emotion recognition, in: 2019 International Conference On Multimodal Interaction, 2019, pp. 562–566.
[27] C. Wang, Y. Ren, N. Zhang, F. Cui, S. Luo, Speech emotion recognition based on multi-feature and multi-lingual fusion, Multimed. Tools. Appl. 81 (4) (2022) 4897–4907.
[28] Cámbara, G., Luque, J., & Farrús, M. (2020). Convolutional speech recognition with pitch and voice quality features. arXiv preprint arXiv:2009.01309.
[29] M. Cappellini, B. Holt, Y.Y. Hsu, Multimodal alignment in telecollaboration: a methodological exploration, System 110 (2022) 102931.
[30] K. Yang, H. Xu, K. Gao, Cm-bert: cross-modal bert for text-audio sentiment analysis, in: Proceedings of the 28th ACM international conference on multimedia, 2020, pp. 521–528.
[31] Devlin, J., Chang, M.W., Lee, K., & Toutanova, K. (2018). Bert: pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv: 1810.04805.
[32] ... & A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, I. Polosukhin, Attention is all you need. Advances in Neural Information Processing Systems, 2017, p. 30.
[33] Okabe, K., Koshinaka, T., & Shinoda, K. (2018). Attentive statistics pooling for deep speaker embedding. arXiv preprint arXiv:1803.10963.
[34] Y. Yu, X. Si, C. Hu, J. Zhang, A review of recurrent neural networks: LSTM cells and network architectures, Neural Comput. 31 (7) (2019) 1235–1270.
[35] ... & C. Busso, M. Bulut, C.C. Lee, A. Kazemzadeh, E. Mower, S. Kim, S.S. Narayanan, IEMOCAP: interactive emotional dyadic motion capture database, Lang. Resour. Eval. 42 (2008) 335–359.
[36] Poria, S., Hazarika, D., Majumder, N., Naik, G., Cambria, E., & Mihalcea, R. (2018). Meld: a multimodal multi-party dataset for emotion recognition in conversations. arXiv preprint arXiv:1810.02508.
[37] K. Kim, S. Park, AOBERT: all-modalities-in-One BERT for multimodal sentiment analysis, Information Fusion 92 (2023) 37–45.
[38] Y. Yi, Y. Tian, C. He, Y. Fan, X. Hu, Y. Xu, DBT: multimodal emotion recognition based on dual-branch transformer, J. Supercomput. 79 (8) (2023) 8611–8633.
[39] N. Majumder, D. Hazarika, A. Gelbukh, et al., Multimodal Sentiment Analysis using Hierarchical Fusion with Context Modeling[J], Knowledge Based Systems 161 (2018) 124–133. DEC.1.

[40] M.R. Makiuchi, K. Uto, K. Shinoda, Multimodal emotion recognition with high-level speech and text features. 2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), IEEE, 2021, pp. 350–357.
[41] J. Santoso, T. Yamada, K. Ishizuka, T. Hashimoto, S. Makino, Speech emotion recognition based on self-attention weight correction for acoustic and text features, IEEe Access. 10 (2022) 115732–115743.
[42] S. Poria, E. Cambria, D. Hazarika, et al., Context-dependent sentiment analysis in user-generated videos, in: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Vancouver, 2017, pp. 873–883.
[43] Ghosal D., Majumder N., Poria S., et al. DialogueGCN: a Graph Convolutional Neural Network for Emotion Recognition in Conversation. 2019.
[44] Hu D., Wei L., Huai X. DialogueCRN: contextual Reasoning Networks for Emotion Recognition in Conversations[J]. 2021.
[45] Hu D., Hou X., Wei L., et al. MM-DFN: multimodal Dynamic Fusion Network for Emotion Recognition in Conversations[J]. arXiv e-prints, 2022.
[46] Sun, Y., Wang, S., Li, Y., Feng, S., Chen, X., Zhang, H., ... & Wu, H. (2019). Ernie: enhanced representation through knowledge integration. arXiv preprint arXiv: 1904.09223.