


Article

Fidgety Speech Emotion Recognition for Learning Process Modeling

Ming Zhu ¹, Chunchieh Wang ² and Chengwei Huang ^{3,*} ¹ School of Information Technology, Yancheng Institute of Technology, Yancheng 224051, China; zhum@ycit.cn² School of Instrument Science and Engineering, Southeast University, Nanjing 210096, China; chunchiehwang@seu.edu.cn³ Zhejiang Laboratory, Hangzhou 310000, China

* Correspondence: huangcwx@126.com

Abstract: In this paper, the recognition of fidgety speech emotion is studied, and real-world speech emotions are collected to enhance emotion recognition in practical scenarios, especially for cognitive tasks. We first focused on eliciting fidgety emotions and data acquisition for general math learning. Students practice mathematics by performing operations, solving problems, and orally responding to questions, all of which are recorded as audio data. Subsequently, the teacher evaluates the accuracy of these mathematical exercises by scoring, which reflects the cognitive outcomes of the students. Secondly, we propose an end-to-end speech emotion model based on a multi-scale one-dimensional (1-D) residual convolutional neural network. Finally, we conducted an experiment to recognize fidgety speech emotions by testing various classifiers, including SVM, LSTM, 1-D CNN, and the proposed multi-scale 1-D CNN. The experimental results show that the classifier we constructed can identify fidgety emotion well. After conducting a thorough analysis of fidgety emotions and their influence on the learning process, a clear relationship between the two was apparent. The automatic recognition of fidgety emotions is valuable for assisting on-line math teaching.

Keywords: speech emotion; AI-assisted teaching; cognitive processes; multi-scale network; fidgety emotion

PACS: 43.72.F; 43.72.L; 43.72.K

MSC: 68T50; 68T45; 68T05



Citation: Zhu, M.; Wang, C.; Huang, C. Fidgety Speech Emotion Recognition for Learning Process Modeling. *Electronics* **2024**, *13*, 146. <https://doi.org/10.3390/electronics13010146>

Academic Editor: Byung-Gyu Kim

Received: 21 November 2023

Revised: 15 December 2023

Accepted: 19 December 2023

Published: 28 December 2023



Copyright: © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

AI-assisted teaching, particularly on online distance learning platforms, gained popularity during and after the COVID-19 pandemic. The automatic recognition of negative emotions is important for studying cognitive outcomes in these educational scenarios.

In past emotion recognition studies, basic types of emotions were extensively studied in controlled and isolated laboratory environments. The relationship between emotion and cognition from a computational perspective has not though been thoroughly explored. Only a limited number of researchers have investigated specific emotions that are associated with the learning and cognitive process.

Pessoa [1] examined emotions from the perspective of brain organization. Pessoa proposed that the conventional categorization of affective and cognitive regions is overly simplified. By emphasizing the intricate interplay between emotion and cognition, this underscores the necessity of obtaining a more comprehensive understanding of how the brain functions and how complex cognitive behaviors emerge. From the computational perspective, Huang et al. [2] conducted studies on the practical problem of speech emotion recognition. They employed various machine learning models to model specific types of emotions bearing on practice, such as confidence, anxiety, and fidgety emotion.

Zepf et al. [3] studied driver's practical emotions, including stress and other types of emotions related to cognitive performance. However, in previous studies, the relationship between emotions and cognitive outcomes has not been fully explored. In particular, the methods suitable for leveraging emotions to aid in cognitive prediction and enhance teaching, as well as how cognitive factors can be used to improve emotion recognition, remain unanswered questions.

Various classification methods have been investigated, including support vector machines, Gaussian mixture models, LSTM (long short-term memory), transformer, and other deep neural networks [4–9]. Shirian et al. [9] investigated a deep graph approach to speech emotion recognition. In their study, they compared various algorithms with graph learning and achieved promising results. However, it is worth noting that only some of the fundamental types of emotions were addressed in their research. Wang et al. [10,11] aimed to investigate the challenges posed by group differences and to sample imbalances in emotion recognition. Their research considered distinctions associated with age and gender. Furthermore, they explored the application of deep neural networks in modeling age and gender differences for speech emotion recognition.

Feature analysis and extraction for emotion recognition have received relatively extensive research attention. Farooq et al. [12] studied feature selection algorithms for speech emotion. In their study, various features were selected and optimized to achieve the best possible results using deep learning. Gat et al. [13] suggested speaker normalization techniques to improve emotion recognition rates. In their study, speaker normalization and self-supervised learning were investigated in detail and experiments were carried out using different databases. Tiwari et al. [14] investigated noisy speech emotion, which is a practical topic that has not been extensively studied. They suggested using data augmentation to improve the modeling of speech emotions. Additionally, they examined a generative noise model for common emotion types. Nevertheless, certain practical emotions were not addressed in their work, and further discussion is needed to explore the practical applications of noisy speech recognition. Lu et al. [15] investigated the generalization of feature extraction, in contrast to domain-specific features, and successfully addressed speaker-dependent issues. The validation of effectiveness was conducted on commonly observed emotion types.

Language-dependent features are also crucial for emotion recognition. Costantini et al. [16] investigated cross-linguistic features of speech emotion. Their study involved the use of various datasets to enable a universal comparison of speech features. Additionally, they explored different machine learning algorithms for emotion modeling and analyzed their generalization capabilities. Saad et al. [17] explored language-independent emotion recognition with a focus on addressing cross-database and cross-language recognition challenges. In their research, they examined and analyzed fundamental speech features, including pitch frequency, formant frequency, and intensity. They also extended their analysis to compare these features between English and Bangla languages. However, there was potential for addressing feature normalization issues further in their work, and the authors did not investigate the relationships between these features, cognition, and personality.

In summary, conventional emotion recognition studies are still limited in methodology, focusing solely on acoustic and computational aspects, while overlooking the intricate relationship between emotion and cognitive processes. Our approach involves employing multi-scale CNNs for modeling and, from a computational perspective, studying the connection between "fidgety" emotion and cognitive processes. We explore how to leverage emotion recognition results to enhance cognitive prediction and improve online teaching.

Fidgety emotion is an important emotional category that differs from traditional emotion research, which focuses on basic emotional categories. Fidgety is a complex emotion with practical value. It holds particular practical significance in the processes of learning and cognition, as it significantly influences cognitive abilities, behavioral control, and psychological stability. While traditional sentiment and emotion recognition (SER)

research extensively cover the six basic emotions, like happiness, anger, surprise, sadness, fear, and disgust, there has been relatively limited research on complex emotions.

2. The Eliciting Experiment and Data Collection

In this section, we introduce our eliciting experiment [18], which involves math problem solving as a cognitive task. During this task, subjects (students) are required to verbally report their outcomes, allowing us to collect speech containing various emotions.

In the Schachter–Singer two-factor theory [19], also known as cognitive arousal theory, it is suggested that emotions are the result of a two-step process. First, individuals experience physiological arousal in response to a stimulus, which can be a general state of physiological excitation. Then, they use cognitive appraisal and external cues to label or interpret that arousal as a specific emotion. According to this theory, the cognitive interpretation is critical in determining which emotion is experienced. Based on the cognitive arousal theory, we make the assumption that the generation of negative emotions, such as feeling fidgety, frustrated, or nervous, may interfere with other cognitive processes, such as math calculations. When a student becomes distracted due to these emotions, it can lead to lower performance in math learning.

Eliciting fidgety emotion using repeated and complex math calculations as an external stimulus aligns with the first factor of the cognitive arousal theory, which involves triggering physiological changes during math tasks. In the following sub-sections, we provide a detailed description of the elicitation and data collection process.

Fidgety emotion is an important practical emotion related to cognition. It often emerges in situations where our minds are engaged, seeking stimulation, or grappling with complex thoughts. This emotion can be manifest in cognitive performance as well as physical behavior. Fidgety emotion can have a range of negative impacts on cognitive functioning and overall well-being. When excessive, it can disrupt one's ability to concentrate and complete tasks efficiently. Persistent fidgeting can be distracting to both the individual and those around them, making it challenging to engage in activities that require sustained attention, such as studying or participating in meetings.

2.1. The Cognitive Task

Cognitive processes include engaging with sequences of mathematical calculation topics. As illustrated in Figure 1, participants in the study undertook cognitive tasks by solving a series of mathematical problems. Throughout this learning process, we captured voice data from the participants using a voice interface. This was utilized, in particular, during repetitive math calculations to elicit fidgety emotions from the participants, enabling the collection of high-quality, naturalistic speech data. We systematically observed and annotated the emotions expressed in each oral report (speech data), while also documenting test scores and individual improvements. Additionally, we recorded the associated mathematical topics as part of the learning history data.

2.2. Data Annotation

Data annotation for emotion recognition necessitates precise emotion labeling across diverse contexts, accounting for factors like cultural nuances, personality, and environmental stimuli. Ensuring inter-annotator agreement through guidelines, training, and regular quality checks is crucial. When selecting data for annotation, diversity is prioritized to train robust models capable of recognizing emotions in various cognitive scenarios and across different demographics.

We employed the Self-Emotion Assessment Scale before and after the math task to monitor emotions. We also conducted a listening test with 12 annotators to label emotions as fidgety, stressed, happy, or neutral. If speech proved challenging to categorize under any of these emotions, we assigned it an "other" label.

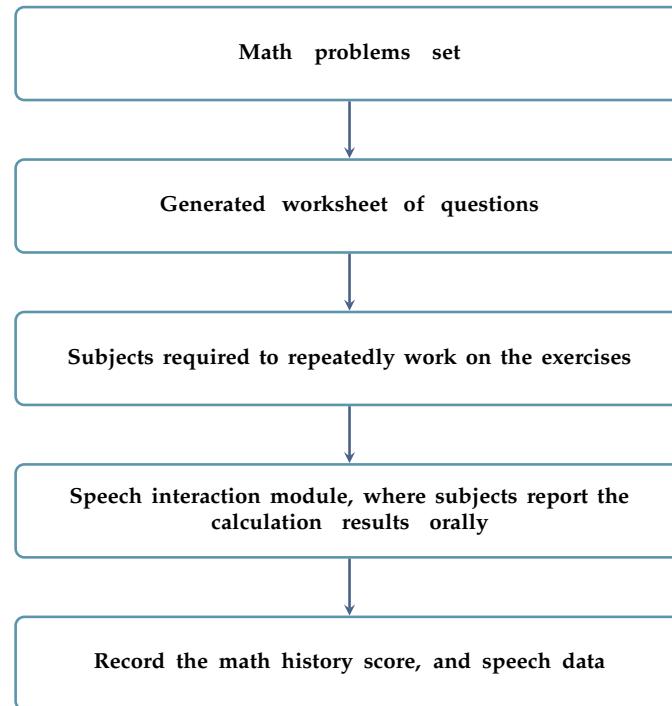


Figure 1. Flowchart of the math exercises and the speech elicitation.

After annotation, consolidating labels from different annotators can be achieved using the analytic hierarchy process (AHP) [20]. This method helps weigh and prioritize the annotations, facilitating assignment of a consensus or aggregated label that reflects the collective judgment of the annotation.

Each emotion annotation divides the intensity of the specific emotion into five levels: 1, 3, 5, 7, and 9. Hence, the comparison matrix is represented as P :

$$P = \begin{bmatrix} 1 & 1/3 & 1/5 & 1/7 & 1/9 \\ 3 & 1 & 1/3 & 1/5 & 1/7 \\ 5 & 3 & 1 & 1/3 & 1/5 \\ 7 & 5 & 3 & 1 & 1/3 \\ 9 & 7 & 5 & 3 & 1 \end{bmatrix} \quad (1)$$

The eigenvalue can be computed as $\lambda_{max} = 5.2375$. The weight vector W is:

$$W = [0.0561, 0.1067, 0.2170, 0.4401, 0.8630]^T \quad (2)$$

The consistency index CI is:

$$CI = \frac{\lambda_{max} - n}{n - 1} = 0.06 \quad (3)$$

The consistency ratio is thus: $CR = CI/RI = 0.06/1.12 = 0.053$, since $CR < 0.1$, it satisfies the consistency requirement.

Finally, we collected a dataset comprising 36 subjects (18 females, 18 males) who volunteered to take part in the data collection, with a total of 4389 annotated emotional speech samples. Among these, there were 1082 labeled as “fidgety”, 858 as “stressed”, 855 as “happy”, 929 as “neutral”, and 665 as “others”.

The distribution of samples is further illustrated in Figure 2. We can see that the utterances had a relatively balanced distribution across different ages and genders. All speakers were Chinese native speakers and the oral test was carried out in standard Chinese.

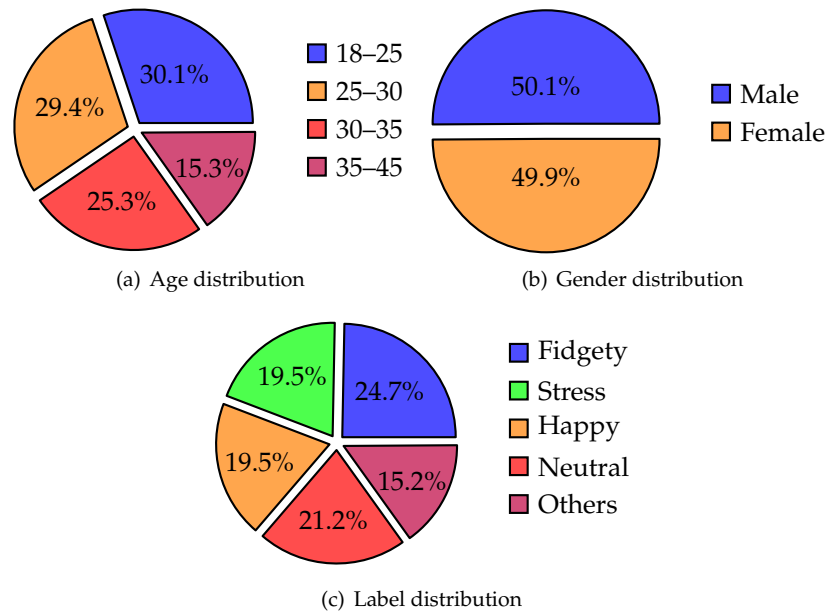


Figure 2. Age, gender, and label distribution among the annotated utterances.

Example of the fidgety emotional speech (female) is shown in Figure 3. The spectrogram and pitch frequency are plotted.

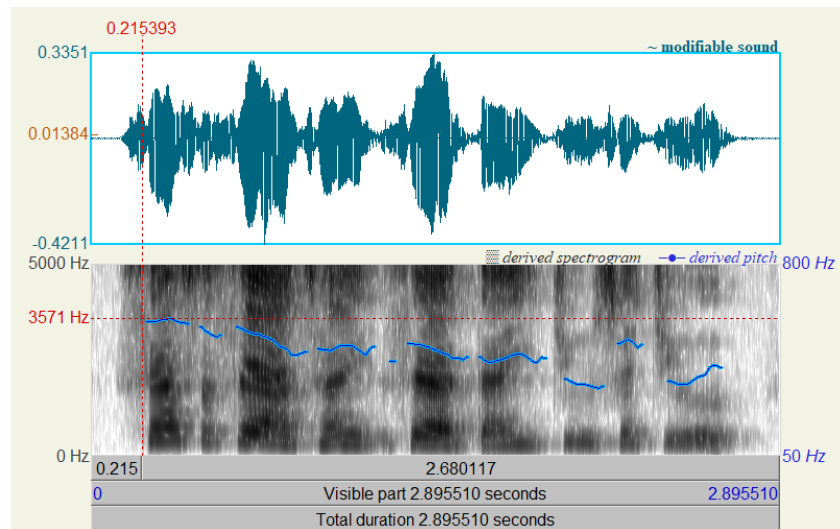


Figure 3. Example of speech signal recording: time–domain waveform, spectrogram, and pitch contour.

3. Methodology

3.1. Multi-Scale CNN for Emotion Recognition

We propose an end-to-end speech emotion model based on a multi-scale one-dimensional (1-D) residual convolutional neural network. The data input to the network is the raw waveform, and the output is the probability corresponding to various emotion categories (including the fidgety emotion).

Multi-scale CNN [21] was used to model and identify the emotional categories. We adopted a time series modeling method to perform 1-D convolution on the scale of the emotional speech signal. We extended the model for application to recognizing fidgety speech emotions. The network architecture is shown in Figure 4.

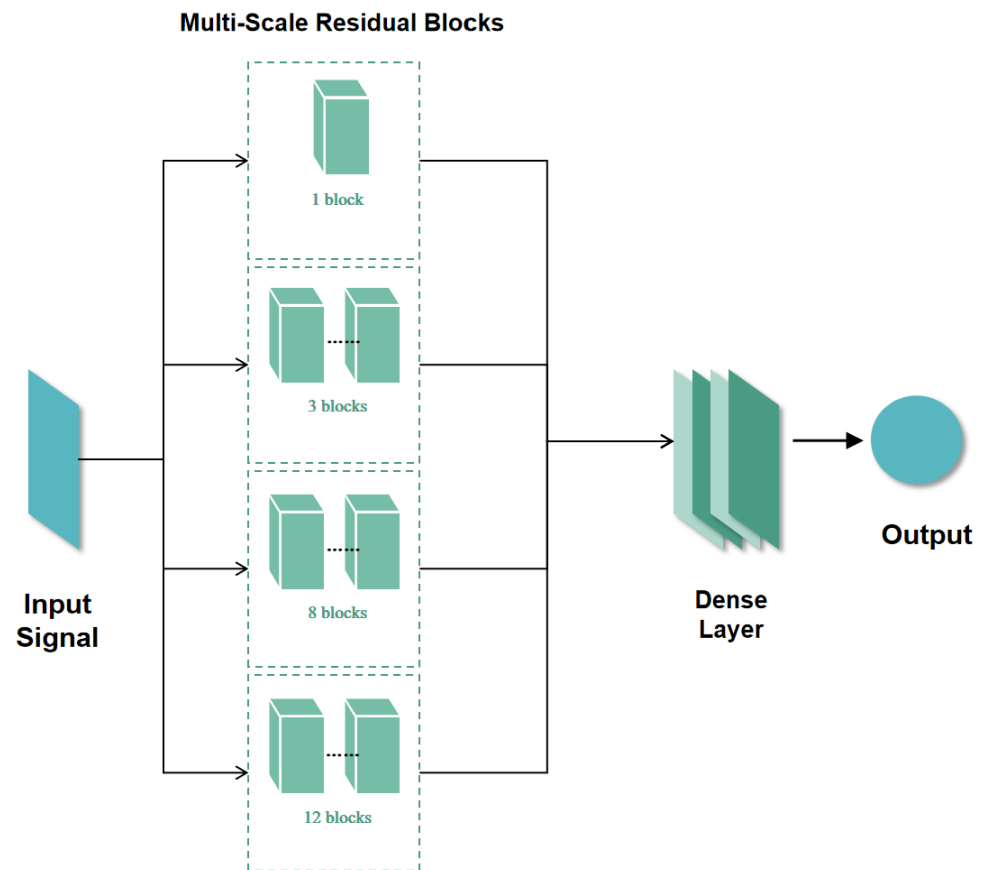


Figure 4. The overall process of multi-scale residual neural network development for emotional feature classification.

The role of dilated convolution is to carry out local feature processing, which is suitable for the representation learning of time-series signals, extracting time-series features through convolution, and is also suitable for modeling sequence data, such as speech emotion.

Given that emotions are expressed over varying durations, and time-domain changes are crucial arousal and valence features, increasing the dilation rate to a value greater than one introduces gaps between the values in the filter. With a larger dilation rate, these gaps become wider, allowing the filter to capture information from a broader receptive field within the emotional speech signal.

Each network block comprises a dilated convolution layer, batch normalization, a residual shortcut connection, and a ReLU layer, all arranged to extract the emotional features from the raw time signal, as shown in Figure 5.

The residual network was proposed by Kaiming He [22]. By introducing a short-cut to avoid problems such as gradient explosion, the network depth can be greatly increased, so that very deep networks can also converge well in training. The residual module is the basic unit that makes up the residual network. Many residual modules cascaded together can improve the effect of representation learning and enable the construction of effective speech emotional features.

In our model, the ReLU function is used as the activation function. We choose 1, 3, 8, or 12 residual blocks for the multi-scale blocks. The optimizer we choose is Adam. The learning rate is set to 0.01, and the loss function is a cross-entropy function.

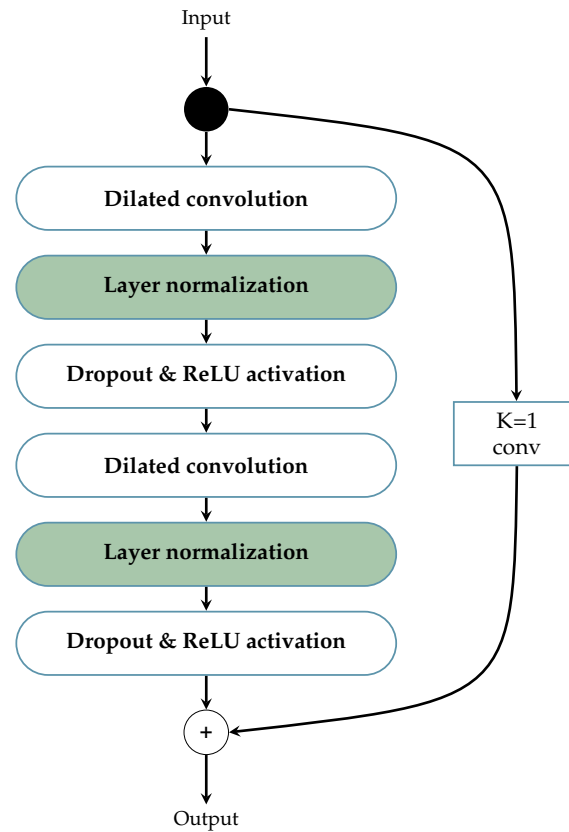


Figure 5. Depiction of the residual neural network block.

3.2. Emotion Recognition and Cognitive Outcome Prediction

As suggested by the Schachter–Singer two-factor theory [19], our eliciting experiments serve as the stimulus to the subjects, provoking physiological changes. Subsequently, during the second cognitive stage, emotions such as fidgety emotions are generated. In our computational model, we assume that both the second stage and the presence of negative emotions will exert an influence on the cognitive outcome.

Through the stimulation of cognitive tasks, which involve mathematical calculations, it is probable that the underlying two-factor process that triggers emotions can also influence cognitive processes. We observe and record changes in cognitive processes from an external perspective, including the problem-solving speed, the question difficulty, and the answer accuracy, which together form a cognitive vector.

As shown in Figure 6, by leveraging these cognitive vectors, we assist in emotion recognition, assuming that there exists a certain relationship between cognitive processes and negative emotions (such as the fidgety emotion). Modeling this probability condition can potentially enhance the results of emotion recognition.

Conversely, based on the outcomes of emotion recognition, as well as the historical data on problem-solving speed and answer accuracy rates, it is possible to predict the probability of correctly answering the next question.

The cognitive vector is defined as a set of metrics shown in Equation (4).

$$\text{Cog_vec} = \{\text{speed}, \text{diff}, \text{rate}\} \quad (4)$$

Speed denotes the measure of the average time spent on one problem (1/time spent), diff denotes the difficult level, and rate denotes the accumulated percentage of correct answers.

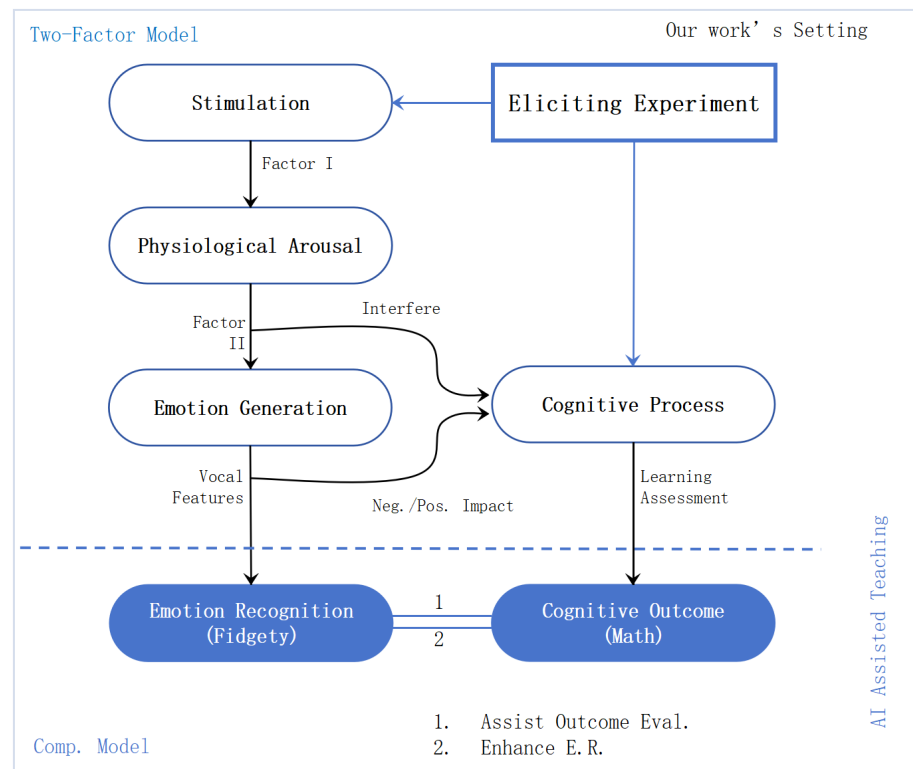


Figure 6. The proposed computational model for cognitive feedback and AI-assisted cognitive outcome prediction.

As depicted in Figure 7, we utilise a cognitive vector to enhance the function of emotion recognition. This cognitive vector is created by incorporating cognitive metrics, specifically, the accuracy of mathematical calculations and the historical score record, as outlined earlier. The resultant “cognitive vector” is subsequently transmitted to the machine learning classifier for emotion recognition, in conjunction with an emotion vector generated from the probability outputs of the residual network. The machine learning algorithm chosen for combining cognitive and emotional data is the Decision Tree, which offers a more comprehensible representation of the relationship between emotional states and cognitive outcomes.

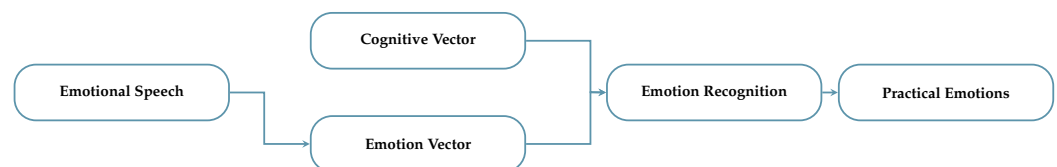


Figure 7. Emotion recognition using cognitive vector.

The machine learning algorithm chosen for integrating cognitive and emotional data is the C5.0 Decision Tree. C5.0 is a sophisticated type of decision tree, renowned for its adaptability and effectiveness in classification and regression tasks. It adeptly partitions the data into subsets by selecting the most informative features, rendering it a valuable tool in our context.

In our specific case, we employ the C5.0 Decision Tree algorithm to amalgamate cognitive and emotional information. It is thoughtfully configured with a maximum depth set at 10 and a requirement for at least five samples to initiate node splitting. In evaluating the quality of these splits, we employ the information gain criterion, a hallmark of C5.0’s advanced decision-making process.

C5.0 Decision Trees are particularly valued for their capacity to unveil the significance of features within a dataset. Given our objective of harmonizing cognitive and emotional data for the recognition of emotional states, understanding which features or metrics exert the most influence becomes paramount. C5.0 decisively illuminates the relative importance of cognitive and emotional metrics, which, in turn, underpins the accuracy of our predictive model.

In this paper, the formulation of the cognitive vector signifies the state observed during the question-solving process. Emotions undeniably exert a noticeable impact on the precision of responses and the pace of question-solving. With this underlying hypothesis, we devised a statistical model to establish statistical connections among the variables, thereby increasing the prediction accuracy. Subsequently, the prediction of the question results' precision serves as a means to corroborate the hypothesis concerning the influence of fidgety emotional states on cognitive speed and cognitive reasoning processes.

$$Cog_vec^{i-1} = \{rate1, rate2, rate3, \dots, speed1, speed2, speed3, \dots\} \quad (5)$$

where i stands for the current index number.

$$Emo_vec = \{emo_vec1, emo_vec2, emo_vec3, \dots\} \quad (6)$$

where $emo_vec = \{p1, p2, p3, \dots\}$ denotes the probability of each emotion type. We focus on the negative emotions, e.g., fidgety emotion, and their impacts on cognitive outcomes.

The input includes the cognitive vectors, emotion vectors, and the cognitive difficulty level, and the output is: $Output = \{Rate, Speed\}$.

As illustrated in Figure 8, in contrast to the emotion recognition process, predicting the cognitive outcomes also involves the outcomes of emotion recognition. When considering the accuracy of mathematical calculations, emotional states play an influential part. In the algorithmic flow presented, we demonstrate the close relation between the emotion category and the preceding cognitive vector, jointly facilitating the prediction of cognitive outcomes, encompassing both the correctness rate, which is the cumulative percentage of correct answers, and the speed of math problem solving.

The predictive model here is Decision Tree. The parameter settings are carefully adjusted, achieving an equilibrium between model complexity and generalization. The maximum depth of the tree is set to 15, enabling the tree to explore the data more comprehensively. The minimum number of samples per node threshold is set to 8, ensuring that nodes split only when a sufficient number of data points are present, thus promoting a more robust and generalized model.

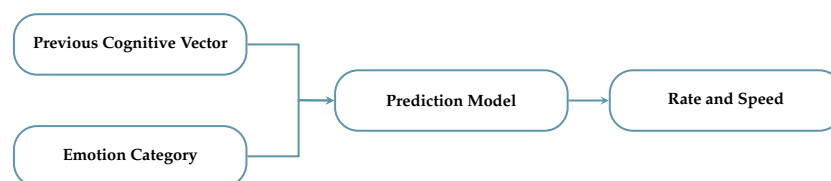


Figure 8. Cognitive vector prediction using emotional states.

4. Experimental Results

The statistics pertaining to the sample distribution within our dataset utilized for this experiment are presented in Table 1. Our dataset comprises a total of 4389 samples, with each mathematical assignment item associated with approximately 5–7 oral report utterances. Furthermore, our dataset contains 665 math assignment questionnaires. The train-validation-test split ratio is set at 7:1:2, resulting in 878 samples allocated for testing. The training samples are randomly selected and mixed; thus, it is speaker independent. The model ability is not dependent on any specific speaker. It has good ability to be generalized to different speakers.

The training of the emotion recognition classifier is a single task. The emotion classifier is trained separately using the emotion labels. The speed and rate for cognitive prediction is estimated independently in the first place, and they can be improved by the emotion recognition results.

Table 1. Dataset sample distribution.

Emotion Type	Sample Size (Total)	Sample Size (Male)	Sample Size (Female)
Fidgety	1082	542	540
Stress	858	430	428
Happy	855	430	425
Neutral	929	466	463
Other	665	333	332

The results for emotion recognition are displayed in Table 2. The confusion matrix highlights the performance of our proposed method, which is built upon a multi-scale 1-D residual network. It is evident from the matrix that fidgety emotion and other cognitive processes are accurately identified.

Table 2. Confusion matrix of fidgety emotion recognition using multi-scale 1-D residual network.

Actual Emotion	Predicted Emotion (%)				
	Fidgety	Stress	Happy	Neutral	Others
Fidgety	85.1	4.5	2.0	3.4	5.0
Stress	5.1	83.4	6.5	2.2	2.8
Happy	2.1	1.5	85.6	4.2	6.6
Neutral	4.3	6.1	3.3	81.3	5.0
Others	3.4	5.3	2.9	7.9	80.5

In order to demonstrate the advantage of our proposed method, we compare it with a basic 1-D convolution model, LSTM (long short-term memory) [23], and SVM. As shown in Figure 9, four emotion classes, fidgety, stress, happy, neutral, and “other” emotion types are modeled and compared. The recognition rates observed show that our proposed multi-scale 1-D residual convolutional network outperformed the rest.

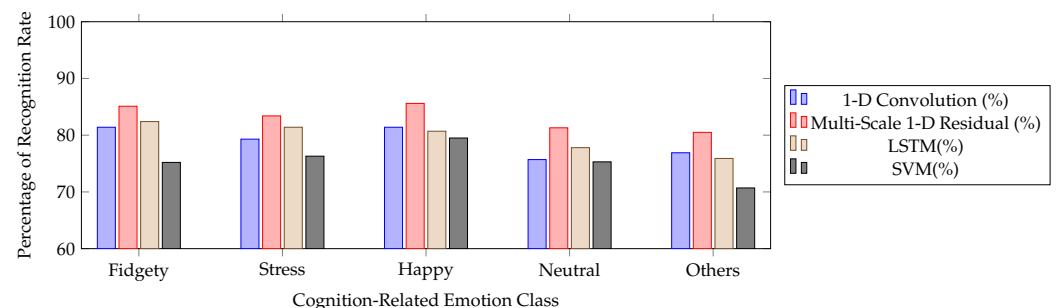


Figure 9. Comparison of algorithms for averaged recognition rates.

Parameter Settings

In order to better compare the different classifiers and to more easily reproduce the models, we describe the parameters used for the basic 1-D convolution model, the LSTM model, and the SVM model. For the basic 1-D convolutional model, identical residual blocks are employed. We maintain a fixed number of residual blocks at three, in contrast to

the multi-scale model where the scale may vary. We further choose the ReLU function as the activation function. For LSTM, we employ the ReLU activation function, the cross-entropy loss function, and set the dropout rate to 0.2. The Adam optimizer is utilized for training the model. For SVM, the radial basis function (RBF) kernel is used as the kernel function, after being compared with the polynomial kernel and the linear kernel.

As shown in Table 3, using the proposed computational model described in the methodology section, we can improve the emotion recognition results by merging the cognitive vector in the recognition process. The results show that fidgety and other cognition-related emotions are improved considerably. The recognition rate for the fidgety emotion is improved from 85.1% to 94.6%.

Table 3. Confusion matrix of fidgety emotion recognition with cognitive vector.

Actual Emotion	Predicted Emotion (%)				
	Fidgety	Stress	Happy	Neutral	Others
Fidgety	94.6	1.5	1.2	1.5	1.2
Stress	0.2	91.3	1.3	3.3	3.9
Happy	0.4	1.6	89.8	3.5	4.7
Neutral	2.8	3.2	3.2	87.3	3.5
Others	1.7	3.6	4.3	5.2	85.2

As depicted in Figure 10, the utilization of cognitive vectors was shown to enhance recognition rates, particularly in the case of negative emotions, like “fidgety”, which exhibited a more pronounced improvement compared to emotions less closely associated with cognitive processes. The results highlight considerable enhancements in emotion recognition rates when cognitive vectors are integrated. Across various emotional categories, the incorporation of cognitive vectors consistently outperforms recognition which relies solely on emotion-related features. Notably, the most substantial improvements were evident in the “Fidgety” and “Stress” categories, where recognition rates increased by 9.5 and 7.9 percentage points, respectively. This suggests that cognitive vectors excel at capturing subtleties in the fidgety emotional state. However, even in the “Happy” and “Neutral” categories, there were noteworthy improvements of 4.2 and 6.0 percentage points, underscoring the versatility of cognitive vectors in enhancing recognition accuracy across a spectrum of emotional classifications.

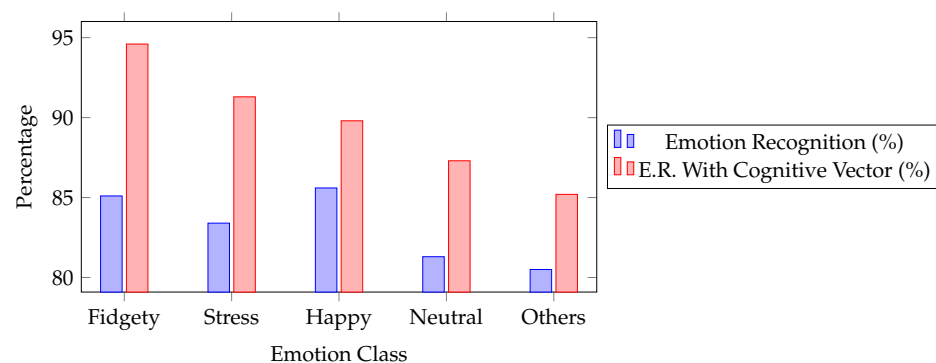


Figure 10. Improvements in recognition rates over emotion classes using cognitive vector.

In our cognitive outcome prediction, we determine the prediction accuracy as the percentage of correct predictions for both right and wrong answers. The math problems’ difficulty levels are categorized into “easy” and “hard”. Along with the difficulty level, different math topics are incorporated as features in our prediction model.

By leveraging peer performance in the math assignment correctness results, we employ an XGBoost classifier to predict future math problem outcomes (as a base prediction model, without considering the emotional states). The model takes as input the current math topics, encoded as one-hot vector IDs, along with the difficulty levels, and the historical assignment results, encompassing both correct and incorrect answers for each past topic, as well as the corresponding time spent on each. The classifier's parameters are configured as follows: 500 for $n_estimators$, 0.01 for $learning_rate$, and 4 for max_depth .

In this experiment, the cognitive prediction results without using emotional information are shown in Table 4. "Speed" denotes the measure of the average time spent on one problem (1/time spent). By using an emotion vector, we can improve the cognitive outcome prediction results. As shown in Table 5, both the rate and speed predictions were improved. We can see that the improvements were considerable when fidgety and stress emotions were present. In the 'easy' category, the rate prediction accuracy improved from 80.1% to 87.7% for the fidgety emotion. In the 'hard' category, the rate prediction accuracy improved from 81.5% to 89.5% for the fidgety emotion. We can conclude that using the emotional states labels contributes to the prediction of cognitive outcomes.

Table 4. Cognitive outcome prediction.

Difficulty Level	Emotional State	Rate Prediction Accuracy	Percentage Error of Speed Prediction
Easy	Fidgety	80.1%	9.3%
Easy	Stress	82.2%	7.8%
Easy	Happy	80.4%	8.1%
Easy	Neutral	84.5%	6.7%
Easy	Others	83.3%	8.7%
Hard	Fidgety	81.5%	9.3%
Hard	Stress	84.4%	7.8%
Hard	Happy	82.4%	10.1%
Hard	Neutral	85.7%	9.7%
Hard	Others	82.4%	7.7%

Table 5. Enhanced cognitive outcome prediction with emotion vector.

Difficulty Level	Emotional State	Rate Prediction Accuracy	Percentage Error of Speed Prediction
Easy	Fidgety	87.7%	6.6%
Easy	Stress	87.5%	5.7%
Easy	Happy	82.4%	6.3%
Easy	Neutral	85.7%	4.4%
Easy	Others	83.9%	7.9%
Hard	Fidgety	89.5%	6.1%
Hard	Stress	88.2%	5.2%
Hard	Happy	86.2%	8.1%
Hard	Neutral	87.4%	8.5%
Hard	Others	83.1%	7.1%

5. Discussion

In our emotion recognition and cognitive prediction experiments, the dataset, composed of 4389 samples, was systematically divided into training, validation, and test sets, with a significant allocation for testing (878 samples). The results for emotion recognition underscore the effectiveness of the proposed multi-scale 1-D residual convolutional network. Notably, the confusion matrix facilitated accurate identification, particularly when discerning fidgety emotion and other cognitive processes.

Comparative analysis with traditional models, such as basic 1-D convolution, LSTM, and SVM, showed the superior performance of the proposed multi-scale 1-D residual network across four emotion classes. The subtle improvements observed, particularly in the recognition of fidgety emotion (94.6% from 85.1%), underscore the model's ability to capture nuanced variations in emotional states.

Furthermore, the integration of cognitive vectors in the emotion recognition process highlights significant enhancement in identifying fidgety and other cognitive-related emo-

tions. These improvements span various emotional categories, with the most substantial gains observed in the recognition of the fidgety emotional state.

The incorporation of emotion vectors in cognitive prediction aligns with the two-factor model, considering physiological arousal and cognitive processes. This approach, attuned to understanding and adapting to students' emotional states during cognitive tasks, offers a nuanced perspective on enhancing the learning experience.

6. Conclusions

In this paper, we present a computational model for emotion recognition and cognitive prediction based on the well-known two-factor model, which takes into account physiological arousal and cognitive processes. We approach the problem of emotion recognition from the perspective that its generation is closely intertwined with cognitive processes. Our methodology began with the creation of an eliciting experiment for collecting emotional speech data during a mathematical cognitive task. Subsequently, we developed a computational model employing a 1-D residual network.

Through comparative analysis among various machine learning classifiers, we established that our proposed approach excels in recognizing emotions with cognitive relevance. Furthermore, we demonstrated the potential utility of emotion recognition in assisting cognitive outcome prediction. This development has promising implications for applications in AI-assisted teaching.

Author Contributions: Conceptualization, C.H.; methodology, M.Z., C.W. and C.H.; software, M.Z., C.W. and C.H.; validation, M.Z. and C.H.; investigation, C.H.; resources, C.H.; writing—original draft, M.Z., C.W. and C.H.; supervision, C.H. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study. Written informed consent has been obtained from the participants to publish this paper.

Data Availability Statement: Data are contained within the article.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Pessoa, L. On the relationship between emotion and cognition. *Nat. Rev. Neurosci.* **2008**, *9*, 148–158. [[CrossRef](#)] [[PubMed](#)]
2. Huang, C.; Jin, Y.; Zhao, Y.; Yu, Y.; Zhao, L. Recognition of practical emotion from elicited speech. In Proceedings of the 2009 First International Conference on Information Science and Engineering, Nanjing, China, 26–28 December 2009; pp. 639–642.
3. Zepf, S.; Hernandez, J.; Schmitt, A.; Minker, W.; Picard, R.W. Driver emotion recognition for intelligent vehicles: A survey. *ACM Comput. Surv. (CSUR)* **2020**, *53*, 1–30. [[CrossRef](#)]
4. Atmaja, B.T.; Sasou, A.; Akagi, M. Survey on bimodal speech emotion recognition from acoustic and linguistic information fusion. *Speech Commun.* **2022**, *140*, 11–28. [[CrossRef](#)]
5. Wagner, J.; Triantafyllopoulos, A.; Wierstorf, H.; Schmitt, M.; Burkhardt, F.; Eyben, F.; Schuller, B.W. Dawn of the transformer era in speech emotion recognition: Closing the valence gap. *IEEE Trans. Pattern Anal. Mach. Intell.* **2023**, *45*, 10745–10759. [[CrossRef](#)] [[PubMed](#)]
6. Wani, T.M.; Gunawan, T.S.; Qadri, S.A.A.; Kartiwi, M.; Ambikairajah, E. A comprehensive review of speech emotion recognition systems. *IEEE Access* **2021**, *9*, 47795–47814. [[CrossRef](#)]
7. Kanwal, S.; Asghar, S. Speech emotion recognition using clustering based GA-optimized feature set. *IEEE Access* **2021**, *9*, 125830–125842. [[CrossRef](#)]
8. Sun, L.; Zou, B.; Fu, S.; Chen, J.; Wang, F. Speech emotion recognition based on DNN-decision tree SVM model. *Speech Commun.* **2019**, *115*, 29–37. [[CrossRef](#)]
9. Shirian, A.; Guha, T. Compact graph architecture for speech emotion recognition. In Proceedings of the ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Toronto, ON, Canada, 6–11 June 2021; pp. 6284–6288.
10. Wang, Z.; Tashev, I. Learning utterance-level representations for speech emotion and age/gender recognition using deep neural networks. In Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP), New Orleans, LA, USA, 5–9 March 2017; Volume 17, pp. 5150–5154.

11. Mustaqeem; Kwon, S. Att-Net: Enhanced emotion recognition system using lightweight self-attention module. *Appl. Soft Comput. J.* **2021**, *102*, 107101. [[CrossRef](#)]
12. Farooq, M.; Hussain, F.; Baloch, N.K.; Raja, F.R.; Yu, H.; Zikria, Y.B. Impact of feature selection algorithm on speech emotion recognition using deep convolutional neural network. *Sensors* **2020**, *20*, 6008. [[CrossRef](#)] [[PubMed](#)]
13. Gat, I.; Aronowitz, H.; Zhu, W.; Morais, E.; Hoory, R. Speaker normalization for self-supervised speech emotion recognition. In Proceedings of the ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Singapore, 23–27 May 2022; pp. 7342–7346.
14. Tiwari, U.; Soni, M.; Chakraborty, R.; Panda, A.; Koppurapu, S.K. Multi-conditioning and data augmentation using generative noise model for speech emotion recognition in noisy conditions. In Proceedings of the ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, 4–8 May 2020; pp. 7194–7198.
15. Lu, C.; Zong, Y.; Zheng, W.; Li, Y.; Tang, C.; Schuller, B.W. Domain invariant feature learning for speaker-independent speech emotion recognition. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2022**, *30*, 2217–2230. [[CrossRef](#)]
16. Costantini, G.; Parada-Cabaleiro, E.; Casali, D.; Cesarini, V. The emotion probe: On the universality of cross-linguistic and cross-gender speech emotion recognition via machine learning. *Sensors* **2022**, *22*, 2461. [[CrossRef](#)] [[PubMed](#)]
17. Saad, H.F.; Mahmud, M.S.; Hasan, P.M.; Farastu, M.; Kabir, M. Is speech emotion recognition language-independent? Analysis of English and Bangla languages using language-independent vocal features. *arXiv* **2021**, arXiv:2111.10776v2.
18. Pascual-Leone, A.; Herpertz, S.C.; Kramer, U. Experimental designs and the ‘emotion stimulus critique’: Hidden problems and potential solutions in the study of emotion. *Psychopathology* **2016**, *49*, 60–68. [[CrossRef](#)] [[PubMed](#)]
19. Ying, L.; Michal, A.; Zhang, J. A Bayesian Drift-Diffusion Model of Schachter-Singer’s Two-Factor Theory of Emotion. In Proceedings of the Annual Meeting of the Cognitive Science Society, Toronto, ON, Canada, 27–30 July 2022; Volume 44.
20. Munier, N.; Hontoria, E. *Uses and Limitations of the AHP Method*; Springer: Berlin/Heidelberg, Germany, 2021.
21. Zhou, G.; Huang, L.; Li, Z.; Tian, H.; Zhang, B.; Fu, M.; Feng, Y.; Huang, C. Intever public database for arcing event detection: feature analysis, benchmark test, and multi-scale CNN application. *IEEE Trans. Instrum. Meas.* **2021**, *70*, 1–15. [[CrossRef](#)]
22. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 26 June–1 July 2016; pp. 770–778.
23. Yu, Y.; Si, X.; Hu, C.; Zhang, J. A review of recurrent neural networks: LSTM cells and network architectures. *Neural Comput.* **2019**, *31*, 1235–1270. [[CrossRef](#)] [[PubMed](#)]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.