Research Article

Research on Feature Fusion Speech Emotion Recognition Technology for Smart Teaching

Shaoyun Zhang¹ and Chao Li²

¹School of Foreign Languages, Anqing Normal University, Anhui Anqing 246003, China
²School of Foreign Languages, Huazhong University of Science & Technology, Wuhan 430074, China

Correspondence should be addressed to Chao Li; leechao@126.com

Received 8 April 2022; Revised 1 May 2022; Accepted 10 May 2022; Published 23 May 2022

Academic Editor: Le Sun

Copyright © 2022 Shaoyun Zhang and Chao Li. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

At present, there are numerous intelligent teaching applications based on various advanced pattern recognition technologies, such as face sign-in, classroom action recognition, student facial expression recognition, and other systems which have been gradually applied to major schools. Speech emotion recognition can analyze the characteristics of current teaching emotions, so as to discover the rules and details of teachers’ grasp of emotions. In order to realize the task of emotion classification in the process of intelligent teaching through machine learning technology, a teaching speech emotion recognition method with multifeature fusion and deep learning is proposed. The proposed method front-end processes the speech spectrum features in a multifeature fusion manner and combines them with an artificial neural network classifier to form the final speech emotion recognition model. First, after speech preprocessing, three features are selected and fused features using a network structure trained by parallel subnetworks. Then, the classifier used is a hybrid neural network classifier combining a convolutional neural network and a recurrent neural network. Finally, 100 open web courses were used to train and test the model. The test results show that the hybrid neural network sound model using multifeature fusion has good teaching speech emotion recognition capability.

1. Introduction

Emotional measurement in classroom teaching has been a pressing issue in instructional psychology. Classroom emotion is also a very important but easily neglected issue in the teaching process, and it is extremely important in promoting the development of students’ quality. At present, most teaching evaluation methods still mainly use people’s subjective judgments and inferences [1–6]. The method of evaluating the quality of classroom teaching through this manual approach suffers from the problems of uneven evaluation benchmarks, too many competent factors, and poor generalization, and it is impossible to come up with more convincing data from an objective perspective [7–9]. At the same time, traditional classroom teaching tends to focus on one or several classrooms only, and it is impossible to dig out the laws of the existence of excellent classrooms from a macroscopic perspective.

In recent years, big data and artificial intelligence technologies have started to develop rapidly in all walks of life. Intelligent teaching based on AI technology has gained wide attention. Technologies such as face sign-in, cloud classroom, intelligent Q&A, and photo search greatly save teachers’ and students’ time and energy and change the traditional classroom teaching mode. In terms of teaching evaluation, artificial intelligence technology can also be used to achieve a more objective evaluation of smart teaching [10–12]. With a large amount of data as support, the evaluation results can be guaranteed to be convincing. Therefore, the use of deep learning technology to achieve teaching evaluation is of great research importance. The teaching emotion as an aspect of teaching evaluation also has a very practical application value.

Teaching emotions have a very big role in the process of intelligent teaching. A good teacher can accurately grasp the teaching emotion, which can improve students’ attention
Mobile Information Systems

while also making them feel happy and satisfied and eventually make them enjoy the course they are taking [13, 14]. However, the complexity of emotion itself has led to a lack of knowledge on how to use emotion in teaching to optimize instruction. Nowadays, the maturity of deep learning algorithms allows us to have a clearer and more objective perception of emotions. By recognizing and analyzing classroom videos through neural networks, we can comprehensively analyze the characteristics and shortcomings of teaching emotions in excellent classrooms from a macro perspective [15–17]. MOOC platforms have a large number of excellent public class videos, which can be made into a more standard classroom speech emotion dataset after professional audio processing.

In order to fully explore the features of teaching emotion in the classroom, so as to achieve the task of classifying teaching emotion, we propose a multifeature fusion and deep learning method for teaching speech emotion recognition. We use the trained neural network to analyze and compare more public classroom videos, so as to summarize the current emotion patterns of excellent courses.

The rest of the paper is organized as follows: In Section 2, the related works are studied in detail, while Section 3 provides the detailed preprocessing of emotional speech signals. Section 4 provides the details of the three feature extraction methods. Section 5 provides an overall architecture for teaching speech emotion recognition. Section 6 provides results and discussion. Finally, the paper is concluded in Section 7.

2. Related Works

In the 1960s, J.S. Bloom, an American psychologist, began to examine the emotion-related aspects of teaching and learning and proposed the influential affective domain and classification of the educational goals system. Bloom stated that the fundamental value of education is to achieve development rather than a challenge. Ketonen et al [18] proposed a theory of classification of educational goals, which can be classified into categories based on emotions. Christelle et al [19] argue that emotions can be seen as a continuum ordered according to a hierarchy and point out that emotions are not only an interpretation of figurative or abstract matters but also a control of the unconscious. These studies overcame the inability to practice traditional methods of evaluating affect in teaching and made it possible to have a concrete operational approach to evaluating affect in teaching.

However, the above method descriptions are too abstract, leading to difficulties in achieving accurate category classification by machines without relevant knowledge. Speech signals are the fastest and most natural way of interaction between humans. Speech emotion recognition is particularly useful for applications that require natural human-computer interaction, such as Web TV, fatigued driving, language translation, and distance learning. This is because the emotional state of the speaker plays a crucial role in all aspects of communication. The main purpose of employing speech emotion recognition is to adjust the system’s functions accordingly when emotions are detected in the voice. Currently, speech emotion recognition systems consist of two phases [20–23]: (1) front-end processing (feature extraction), which extracts the appropriate available speech features from the available speech data. Each feature parameter in speech emotion recognition contains only part of the information of the speech signal. In order to fully characterize the speech signal, the combined use of various feature parameters becomes an important research direction. How to combine, transform, and trade-off various feature parameters is a problem that must be solved; (2) classifier, which determines the potential emotion of speech. Most of the current research has focused on the selection of classifiers. The main types of various existing classifiers are hidden Markov models (HMM), Gaussian mixture models (GMM), artificial neural networks (ANN), and support vector machines (SVM) [24–26]. These classifiers are widely used in speech emotion recognition in their individual form or in a combined form.

To address the fusion problem of feature parameters and the selection problem of classifiers, a multifeature fusion approach is proposed for front-end processing of speech spectrum features and combining them with an artificial neural network classifier to form a final speech emotion recognition model. The classifier used is a hybrid neural network (HNN) classifier combining convolutional neural network (CNN) and recurrent neural network (RNN), and the effectiveness of the method is demonstrated experimentally.

The main innovations in this paper are reflected in the following two points: (1) three features are selected based on the similarity and complementarity among sound features, and the features are fused using a network structure trained by parallel subnetworks; (2) in the proposed HNN classifier, the CNN part uses DenseNet with excellent performance in picture recognition tasks, and the RNN part uses the DenseNet with excellent performance in speech recognition and text annotation LSTM neural network, which is commonly used in speech recognition and text annotation tasks. The combination of the two makes it easier to find patterns in data with time-series features and thus identify the target sentiment class more smoothly.

3. Preprocessing of Emotional Speech Signals

The preprocessing of speech signal includes preemphasis, short-time analysis, framing, windowing, and endpoint detection.

(1) Preemphasis: The speech signal spectrum is processed by Fourier transform. The main content of the preemphasis is the high-frequency part, which can ensure the signal spectrum is smoother, thus ensuring its analysis or the analysis of channel parameters is easier.

(2) Short-time analysis: As a whole, the speech signal changes with time, which belongs to a nonstationary process, and the digital signal processing technology for stationary signals cannot be used to process it.
However, if we look at a short-time range (generally considered to be 10–30 ms), the frequency of speech during oral muscle movement is extremely slow. Therefore, the analysis and processing of speech are all based on short time.

(3) Frame splitting: Because in the short time analysis of speech, the signal needs to be divided and processed. A frame is usually 10–30 ms, and in order to achieve a smooth transition and maintain continuity, the adjacent frames are usually overlapped. According to the frame length and step length, all speech frames of a speech segment can be obtained.

(4) Windowing: After framing the speech signal, it is often necessary to use Fourier variation, but then a Gibbs phenomenon occurs. Therefore, we need to add window processing. The commonly used window function is a rectangular window and Hamming window. The Hamming window function is defined as follows:

$$w(n) = 0.54 - 0.46 \cos\left(\frac{2\pi n}{N}\right), \quad 0 \leq n \leq N,$$

where $N$ is the length of the Hamming window.

(5) Endpoint detection: The start and endpoints of the speech signal are accurately found from a segment of speech, and the effective speech signal is separated from the invalid noise signal.

4. Three Feature Extraction Methods

Speech expresses emotion because it contains parameters that characterize emotion. The variation of emotion is reflected by the difference in feature parameters. The focus of feature extraction is to retain information with cause-effect relationships and eliminate information that is not involved with each other. The main features commonly used in speech recognition are sound spectrogram, Meier cepstral coefficient (MFCC) features, and filter bank (FBank) features. In this paper, these three features are selected for teaching speech emotion recognition.

4.1. Sound Spectrogram Characteristics. The speech signal is a one-dimensional signal, and intuitively only the information of speech in the time domain can be seen, but not its frequency domain information. Therefore, this chapter extracts the sound spectrogram features, and the flow chart of extracting sound spectrogram features is shown in Figure 1.

The frame length is set to 25 ms, the frame shift is set to 10 ms, and then, a window is added to each frame, and the Hamming window is used in this paper. The short-term Fourier transform (STFT) is applied to each frame after the windowing. The STFT is introduced to preserve the time-frequency relationship of the speech signal, and the transformation equation is as follows:

$$STFT_x(n, m) = \sum_{k=\infty}^{\infty} z(k) * w(kT - nT)e^{-j2\pi(mF)k},$$

where $z(m)$ is the source signal, $w(m)$ is the window function, $F$ is the scaling factor, and $T$ is the time period.

Finally, each frame of the signal obtained by the transformation is stacked from another dimension to obtain the sound spectrogram features.

4.2. FBank Features. FBank is a feature similar to the way the human ear processes audio, which can optimize the recognition effect of the speech recognition system. In order to maximize the information of the sound signal and get the best feature parameters, the FBank features need to be extracted by splitting frames and adding windows, etc. The flow chart is shown in Figure 2.

Preaggravation refers to the use of the high-pass filter to enhance the high-frequency region of the speech signal and keep it in the whole frequency range from low frequency to high frequency. The high-pass filter function selected for the preaggravation process is

$$y(n) = x(n) - a * x(n - 1),$$

where $x(n)$ is the input signal, $y(n)$ is the output signal, and $a$ is the preemphasis factor. $a$ is a value between 0.9 and 1.0. In this paper, $a = 0.97$. When extracting FBank features, the

---

**Figure 1:** Extraction process of sound spectrogram features.

**Figure 2:** Flowchart of FBank feature extraction.
length of each frame is set to 25 ms, the length of the overlap between the two frames is set to 10 ms, and the window function is selected as the Hamming window. After adding the window function, in order to obtain the spectral energy distribution of the signal, the processed frame signal needs to be subjected to fast Fourier transform (FFT) in the following way:

\[
S_i(k) = \sum_{n=1}^{N} s_i(n)e^{-j2\pi kn/N} \quad 1 \leq k \leq N,
\]

where \(N\) is the number of samples and \(s_i(n)\) denotes the \(i\)-th frame of the input signal.

Since the FBank features take into account the auditory characteristics of the human ear, it is also necessary to transform the obtained frequency domain features into the nonlinear spectrum in the Mel domain. The equation for the Mel transformation is as follows:

\[
\text{Mel}(f) = 2595 \times \log \left(1 + \frac{f}{700}\right),
\]

where \(f\) is the signal frequency.

Set the number of triangular filter banks to \(M\) and set \(f(m), m = 1, 2, \ldots, M\) to be the center frequency of this filter bank. These frequencies are equally distributed on the Mel frequency axis. The frequency-domain features are filtered through the Mel filter bank to obtain the energy value of the corresponding frequency band for each frame of the signal on this filter. Then, logarithmic operations are performed on the resulting energy values to finally obtain the FBank features. The logarithmic energy output of each filter bank is calculated as follows:

\[
s(m) = \ln \left(\sum_{k=0}^{N-1} |S(k)|^2 H(k)\right) \quad 0 \leq m \leq M.
\]

4.3. MFCC Features. Although the FBank feature has closely matched the response characteristics of the human ear, it still has some shortcomings, such as the superposition part between each filter bank of the FBank feature. Therefore, MFCC features have been proposed. The MFCC features can be obtained by performing a discrete cosine transform (DCT) on the base of the extracted FBank features, and the extraction process of MFCC features is shown in Figure 3.

The DCT serves to remove the correlation between the signals in each dimension and maps the signals to a multidimensional space. The discrete cosine transform equation is as follows:

\[
c(n) = \sum_{m=1}^{N} s(m)\cos\left(\frac{mM}{M}(M - 0.5)\right) \quad n = 1, 2, \ldots, L,
\]

where \(L\) is the order of the MFCC coefficient and \(M\) is the number of triangular filters.

In general, to better reflect the dynamic continuity of the signal, the differential form of the static features can be used to represent its dynamic continuity.

\[
\Delta_t = \begin{cases} 
C_{t+1} - C_t, & t < K, \\
\sum_{k=1}^{K} (C_{t+k} - C_{t-k}) / \sqrt{2 \sum_{k=1}^{K} k^2}, & \text{others}, \\
C_{t} - C_{t-1}, & t \geq M - k,
\end{cases}
\]

where \(\Delta_t\) denotes the first-order difference equation. The related parameters of the second-order difference formula can be obtained by substituting the result of the first-order difference into formula (8).

5. Teaching Speech Emotion Recognition Based on Multifeature Fusion and HNN

5.1. Structure of HNN Classifier. The overall architecture of instructional speech emotion recognition consists of three parts, the first two of which constitute the HNN. The first part is a convolutional feature extractor, which takes as input a spectrogram image representation of an audio file. This feature extractor uses DenseNet [27] to convolve the input image and merge it in several steps and generate a spreading feature map. For the feature map represented as \([s(1), s(2), \ldots, s(T)]\) of a given vocalization, where \(s(t)\) is...
the segmented spectrogram of the original audio input and $T$ is the number of segments. We obtain the sequence of feature vectors $[c(1), c(2), \ldots, c(T)]$ after the CNN module.

The second part is a recurrent neural network (RNN) [28]. RNN can handle inputs of various lengths and therefore does not require audio input clips or padding. An LSTM structure was chosen for the RNN that addresses the long-term dependencies present in sequential data. The LSTM learns in a sequence of feature vectors $[c(1), c(2), \ldots, c(T)]$ of the CNN and outputs a sequence $[r(1), r(2), \ldots, r(T)]$. Then, the statistics of the LSTM output are computed through the pooling layer. In the usual case, only the average pooling is applied. In order to obtain richer statistical information about the output of the LSTM network, we perform maximum pooling and minimum pooling. The LSTM is set to have 128 units. The merging process can be expressed as follows:

$$P_{\text{max}} = \{p_{\text{max}}^1, \ldots, p_{\text{max}}^{128}\}, p_{\text{max}}^i = \max_{1 \leq t \leq T} r(t)^i,$$

$$P_{\text{ave}} = \sum_{1 \leq t \leq T} \frac{r(t)^i}{T},$$

$$P_{\text{min}} = \{p_{\text{min}}^1, \ldots, p_{\text{min}}^{128}\}, p_{\text{min}}^i = \min_{1 \leq t \leq T} r(t)^i,$$

where $r(t)^i$ represents the $i$-th element of $r(t)$. The network structure of HNN is shown in Figure 4.

5.2. Multifeature Sound Model Training. In this paper, the network structure of parallel subnetwork training is used to fuse the features. The advantage of feature fusion is that it can obtain the most effective feature and the minimum dimension feature vector set which is beneficial to the final decision. The structure diagram of the parallel subnetwork training is shown in Figure 5.

The feature fusion process is trained using parallel, separate networks for the three features. Each individual network consists of an HNN and performs deep processing of the three different features. A fully connected layer is connected afterwards, whose role is to converge the outputs of the independent networks using a tandem connection to form a sound model. The third part of the overall architecture for teaching speech emotion recognition consists of two fully connected layers and a softmax layer for predicting emotion categories with sizes of 128, 32, and 9, respectively.

Since three different types of features are selected in this paper and their dimensions are different, the traditional Euclidean distance cannot be used for direct comparison. Therefore, this paper introduces the feature proximity and...
uses the average dimensional spacing between vectors instead of Euclidean distance to represent the proximity between different features. The average dimensional spacing between vectors is shown as follows:

\[
\begin{align*}
  d_\mu(i, j) &= \frac{1}{N_i N_j} \sqrt{\sum_{i} \sum_{j} (u_i - u_j)^2}, \\
  d_{\sigma^2}(i, j) &= \frac{1}{N_i N_j} \sum_{i} \sum_{j} (v_i - v_j)^2,
\end{align*}
\]

where \(d_\mu(i, j)\), \(d_{\sigma^2}(i, j)\) denote the mean interval and variance interval of the average dimension among the feature vectors, respectively. \(u_i, u_j\) denote the vectors of the mean values of the class \(i\) and \(j\) sound features, respectively. \(v_i, v_j\) denote the vectors of the variance of the class \(i\) and \(j\) sound features, respectively.

6. Experiment and Result Analysis

6.1. Experimental Data Sources and Preprocessing. The source of the dataset is 100 courses on the MOOC (MUOC) platform of Chinese universities (https://www.cmooc.com). In this paper, we use BeautifulSoup4 crawler based on Python language to extract the video connection of the web page and then use the Requests library to send Post Request to this connection, so as to get the download link of the video. The audio splitting uses the split_on_silence function in the pydub library. Audio splitting determines the speaker’s speech interval based on the silence time, thus splitting the audio of a class into a large number of segments based on the speech. The 100 courses crawled were divided into more than 30,000 audio segments after audio cutting, and manually labeled. The distribution of the labeled data is shown in Figure 6, with 50 classrooms as the training dataset.

It can be seen that there is a serious data imbalance in the labeled teaching emotion dataset. The highest number of “calm” tags was 10,459, accounting for 71.52% of the total data set, while the lowest number of “surprised” tags was only 2, less than 0.1% of the total data set.
Since the distribution of the data set is very unbalanced, we need to take some data preprocessing to ensure that our neural network has some generalization ability so that no serious overfitting will occur. Undersampling is performed on the “Calmness” labels, which account for more than 71% of the total data. The undersampling operation is an operation that discards a large amount of data, and we randomly discarded about 90% of the data volume, leaving a final sample of 1085 “Calmness” labels. Similarly, for smaller data sets, such as the “Tension,” “Hesitation,” and “Satisfaction” labels, we perform an “oversampling” operation. Oversampling is the process of repeating a positive proportion of the data to ensure that the amount of data is not too small to cause overfitting problems. Although the proportion of data may have changed after oversampling and undersampling, the balance of data is better improved. The distribution of the preprocessed teaching emotion data is shown in Table 1.

6.2. Analysis of Emotion Recognition Results. In order to verify the recognition effect of multifeature fusion, this paper forms a comparison experiment by fusing different feature combinations. Because the dimension of spectrogram features is relatively large, its network structure adds a pool layer on the basis of convolution pool convolution to reduce the dimension. The network parameters of HNN are set as shown in Table 2. The error rate comparison results of the test set are shown in Figure 7.

As can be seen from Figure 7, Test I only trained on FBank features for recognition, so Test I had the worst recognition effect, with a recognition error rate of 25.92%. Test II is based on the introduction of MFCC features on top of FBank feature training, and the recognition effect is slightly improved, and the recognition error rate is 0.41% lower than that of Test I. Test III introduces the sound spectrogram feature on the basis of Test I. The recognition effect is also improved, and the recognition error rate is reduced by 0.86% compared with Test I. Test VI introduces both MFCC and sound spectrogram features on the basis of Test I. Analyzing the test results, we can get that the recognition effect of test VI is the best, and its recognition error rate is reduced by 1.28%, 0.87%, and 0.42%, respectively compared with the first three tests. From the overall recognition results, it can be obtained that the HNN sound model with multifeature fusion has better teaching speech emotion recognition ability.

6.3. Teaching Emotion Distribution Pattern. In this paper, based on the data obtained, the classes were classified into the following six categories according to their
emotional characteristics: calm classroom, inquiry classroom, question-and-answer classroom, encouraging classroom, exciting classroom, and balanced classroom. After the identification of teaching emotions, the distribution of categories in 100 courses is shown in Table 3.

Almost all classrooms are dominated by one emotion: the “calm” emotion. In other words, almost all classrooms are currently teacher-led classrooms. Only three inquiry-based classes, in which the percentage of teacher-led lessons is below 50%, are truly student-led. Teachers should explore new ways to guide students and really make them the masters of the classroom.

7. Conclusions

In this paper, we fuse FBank features, MFCC features, and sound spectrogram features for HNN network training, thus building a multifeature sound model. The fusion process uses independent parallel-style subnetwork training, followed by common training through fully connected layers. The HNN classifier used is a hybrid classifier combining CNN and RNN. It is concluded from the experimental results that the recognition effect of the model is optimized with the increase of fused features. For training the teaching speech emotion recognition model, we can achieve fast recognition and classification of emotions in the intelligent teaching process. However, the study has certain shortcomings and needs to be improved: (1) the data set produced has too much imbalance in the data content, which is very unfavorable for neural network training; (2) only 50 classroom test data sets were produced, but they are still insufficient for large-scale training. Both the data and the neural network model have a lot of room for improvement and modification.

Data Availability

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

Conflicts of Interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

References

[1] B. Wang, J. Wang, and G. Hu, “College English classroom teaching evaluation based on particle swarm optimization - extreme learning machine model,” International Journal of Emerging Technologies in Learning (iJET), vol. 12, no. 05, pp. 82–92, 2017.
[2] E. Dehon, E. Robertson, M. Barnard, J. Gunalda, and M. Puskarich, “Development of a clinical teaching evaluation and feedback tool for faculty,” Western Journal of Emergency Medicine, vol. 20, no. 1, pp. 50–57, 2018.
[3] X. Xu, Y. Wang, and S. Yu, “Teaching performance evaluation in smart campus,” IEEE Access, vol. 6, pp. 77754–77766, 2018.
[4] W. A. Spivey and D. F. Caldwell, “Improving mba teaching evaluation: insights from critical incident methodology,” Journal of Marketing Education, vol. 4, no. 1, pp. 25–30, 2016.
[5] X. Xia and J. Yan, “Construction of music teaching evaluation model based on weighted naïve bayes,” Scientific Programming, vol. 2021, pp. 1–9, Article ID 7196917, 2021.
[6] C. Zhang and Y. Guo, “Retracted article: mountain rainfall estimation and online English teaching evaluation based on RBF neural network,” Arabian Journal of Geosciences, vol. 14, no. 17, pp. 1736–2139, 2021.
[7] T. Xia, “Based on big data college physical education teaching evaluation system research,” Journal of Physics: Conference Series, vol. 1744, no. 3, Article ID 032010, 2020.
[8] M. Kazi, “Instructional leadership: teaching evaluation as a key element for 6th grade student’s achievement in mathematics,” International Journal of Educational Management, vol. 35, no. 6, pp. 1191–1204, 2021.
[9] W. Xiao, “Knowledge structure of teaching evaluation research in Chinese universities in the past ten years-an analysis based on CiteSpace map,” Advances in Education, vol. 11, no. 01, pp. 120–128, 2021.
[10] J. Guo, L. Bai, Z. Yu, Z. Zhao, and B. Wan, “An ai-application-oriented in-class teaching evaluation model by using statistical modeling and ensemble learning,” Sensors, vol. 21, no. 1, pp. 241–252, 2021.
[11] X. Zhang, X. Yang, and J. Yang, “Teaching evaluation algorithm based on grey relational analysis,” Complexity, vol. 2021, no. 1, pp. 1–9, Article ID 5596518, 2021.
[12] J. Hu, “Teaching evaluation system by use of machine learning and artificial intelligence methods,” International Journal of Emerging Technologies in Learning (iJET), vol. 16, no. 05, pp. 87–96, 2021.
[13] J. Zhang, “Research on classroom teaching evaluation and instruction system based on gis mobile terminal,” Mobile Information Systems, vol. 2021, pp. 1–11, Article ID 9790766, 2021.
[14] Y. Zheng and H. Ke, “The adoption of scale space hierarchical cluster analysis algorithm in the classification of rock-climbing teaching evaluation system,” Journal of Ambient Intelligence and Humanized Computing, vol. 12, pp. 92–116, 2020.
[15] B. J. Abbaschian, D. Sierra-Sosa, and A. Elmaghraby, “Deep learning techniques for speech emotion recognition, from databases to models,” Sensors, vol. 21, no. 4, pp. 1249–2201, 2021.
[16] N. T. Pham, D. N. M. Dang, and S. D. Nguyen, “A method upon deep learning for speech emotion recognition,” Journal of Advanced Engineering and Computation, vol. 4, no. 4, pp. 273–279, 2020.
[17] B. Kratzwald, S. Illié, M. Kraus, S. Feuerriegel, and H. Prendinger, “Deep learning for affective computing: text-based emotion recognition in decision support,” Decision Support Systems, vol. 115, no. 11, pp. 24–35, 2018.
[18] E. E. Ketonen, J. Dietrich, J. Moeller, K. Salmela-Aro, and K. Lonka, “The role of daily autonomous and controlled educational goals in students’ academic emotion states: an
experience sampling method approach,” *Learning and Instruction*, vol. 53, pp. 10–20, 2018.

[19] C. Declercq, P. Marlé, and R. Pochon, “Emotion word comprehension in children aged 4-7 years,” *The Educational and Developmental Psychologist*, vol. 36, no. 2, pp. 82–87, 2019.

[20] B. W. Schuller, “Speech emotion recognition,” *Communications of the ACM*, vol. 61, no. 5, pp. 90–99, 2018.

[21] N. Vrysas, R. Kotsakis, A. Liatsou, C. Dimoulas, and G. Kalliris, “Speech emotion recognition for performance interaction,” *Journal of the Audio Engineering Society*, vol. 66, no. 6, pp. 457–467, 2018.

[22] M. F. Alghifari, T. S. Gunawan, and M. Kartiwi, “Speech emotion recognition using deep feedforward neural network,” *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 10, no. 2, pp. 554–561, 2018.

[23] Z.-T. Liu, Q. Xie, M. Wu, W.-H. Cao, Y. Mei, and J.-W. Mao, “Speech emotion recognition based on an improved brain emotion learning model,” *Neurocomputing*, vol. 309, no. 12, pp. 145–156, 2018.

[24] A. Hojjatian, D. Taylor, N. Daneshparvar, P. Fagnant, K. Trybus, and K. Taylor, “Cryo-em structure of f-actin decorated by hmm in rigor state,” *Microscopy and Microanalysis*, vol. 27, no. S1, pp. 836–839, 2021.

[25] R. B. Lanjewar, S. Mathurkar, and N. Patel, “Implementation and comparison of speech emotion recognition system using Gaussian mixture model (gmm) and k- nearest neighbor (k-nn) techniques,” *Procedia Computer Science*, vol. 49, pp. 50–57, 2015.

[26] Q. Hou, C. Li, M. Kang, and X. Zhao, “Intelligent model for speech recognition based on svm: a case study on English language,” *Journal of Intelligent and Fuzzy Systems*, vol. 40, no. 2, pp. 2721–2731, 2021.

[27] X. Gao, T. Chen, R. Niu, and A. Plaza, “Recognition and mapping of landslide using a fully convolutional densenet and influencing factors,” *Ieee Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 14, pp. 7881–7894, 2021.

[28] B. Li and Y. He, “A feature-extraction-based lightweight convolutional and recurrent neural networks adaptive computing model for container terminal liner handling volume forecasting,” *Discrete Dynamics in Nature and Society*, vol. 2021, pp. 1–17, Article ID 6721564, 2021.