INDERSCIENCE PUBLISHERS
Linking academia, business and industry through research

International Journal of Continuing Engineering Education and Life-Long Learning

ISSN online: 1741-5055 - ISSN print: 1560-4624
https://www.inderscience.com/ijceell

Classroom student emotion recognition using an improved segmentation clustering and multi-feature fusion emotion recognition algorithm

Xiaohong Wang

DOI: 10.1504/IJCEELL.2026.10079006

Article History:

Received:	15 September 2025
Last revised:	02 February 2026
Accepted:	04 February 2026
Published online:	16 June 2026

Classroom student emotion recognition using an improved segmentation clustering and multi-feature fusion emotion recognition algorithm

Xiaohong Wang

School of Humanities,
Weinan Normal University,
Weinan, 714099, China
Email: xiaohongwx@outlook.com

Abstract: To address noise, speech masking, and weak robustness in classroom emotion recognition, this study proposes a model combining enhanced segmentation clustering and multi-feature fusion. An improved U-Net with local loss supervision first performs denoising. Secondly, using MFCC features combined with Bayesian segmentation and K-means clustering to process speech signals. Finally, MFCC, formant, and pitch features are integrated into an attention-based BiLSTM for emotion recognition. Results show the U-Net achieved a loss of 0.26 after 16 iterations, with PESQ at 3.01 and STOI 84.21%. Segmentation false negative and positive rates were 13.84% and 12.52%. K-means purity reached 90.85%. The multi-feature model attained 92.98% accuracy for excited emotion, and the full system reached 93.62% test accuracy. The model improves recognition in complex classrooms, supporting personalised smart education.

Keywords: emotion recognition; speech signal processing; segmentation and clustering; multi-feature fusion; MFF; support vector machines; SVMs; classroom engagement; deep learning; attention mechanism.

Reference to this paper should be made as follows: Wang, X. (2026) 'Classroom student emotion recognition using an improved segmentation clustering and multi-feature fusion emotion recognition algorithm', *Int. J. Continuing Engineering Education and Life-Long Learning*, Vol. 36, No. 10, pp.1–26.

Biographical notes: Xiaohong Wang received her Bachelor of Arts degree from Shaanxi Normal University, China in 2002, and obtained her Master's in Linguistics and Applied Linguistics from Huazhong University of Science and Technology in 2008. Her research areas cover pedagogy, artificial intelligence, etc. She is currently an associate professor in the School of Humanities, Weinan Normal University, and has been appointed as a social science expert in Weinan City, Shaanxi Province, China. She has published papers in a number of domestic and international journals. Her research interests include curriculum and instruction, and artificial intelligence.

1 Overview

In modern educational philosophy, classroom instruction is no longer merely a one-way transmission of knowledge, but rather a dynamic process where teachers and students jointly construct knowledge, stimulate thinking, and cultivate abilities (Ezquerra et al.,

2025; Aldhilan and Rafiq, 2025). Student engagement and emotional states in the classroom serve as valuable indicators of learning commitment and teaching effectiveness, playing a crucial role in enhancing overall instructional quality. However, current mainstream classroom evaluation methods still face significant challenges (Bie et al., 2024; Tang et al., 2025; Zhang et al., 2023). With the advancement of artificial intelligence technology, intelligent assessment solutions based on computer vision or physiological sensing have emerged. While these methods have enhanced the objectivity of evaluations to some extent, their limitations are becoming increasingly apparent (Zhu et al., 2024; Ashok Kumar et al., 2023). In contrast, emotion recognition (ER) technology based on voice signals offers a significant solution for achieving more precise classroom evaluation. This is due to its inherent advantages in privacy protection and its direct correlation with student engagement.

In recent years, scholars both domestically and internationally have achieved fruitful research outcomes in the field of speech emotion recognition (SER). Falahzadeh et al. (2023) discovered that speech signals are difficult to directly input into deep convolutional neural networks (CNNs) for SER. To address this, they proposed a pre-trained deep CNN model based on chaos maps. Results indicated that this model demonstrated superior recognition performance across multiple public emotion speech datasets (Falahzadeh et al., 2023). Subramanian et al. proposed a novel SER framework to address the challenge of feature selection in regional language ER in India. The method's remarkable 97.96% recognition accuracy on the Tamil emotion dataset was shown by the results (Subramanian and Aruchamy, 2024). Chen et al. (2024) addressed the issue of insufficient effectiveness in general large-scale pre-trained models for SER tasks by proposing an improved emotion-specific pre-trained encoder named Vesper. Results demonstrated that Vesper significantly outperformed other methods in SER across multiple public datasets (Chen et al., 2024). Mishra et al. (2024) used entropy characteristics based on Mel-frequency cepstral coefficients (MFCCs) to improve performance in an attempt to tackle the problem of increasing classification accuracy in SER. Results demonstrated that this approach achieved a classification accuracy of 87.48% in SER (Mishra et al., 2024). Panda et al. (2023) proposed a feature fusion technique to achieve effective feature selection for speech emotion state recognition. Results demonstrated that the proposed classifier achieved a recognition rate as high as 99.64% (Panda et al., 2023). Patnaik (2023) addressed the issues of high computational complexity and strong subjectivity in SER by proposing a classification technique that combines deep sequence long short-term memory neural network with enhanced MFCC characteristics. The robustness and efficacy of improved MFCC as an emotion representation feature were validated by the results, which showed that this method attained up to 98.5% accuracy across six emotion tests (Patnaik, 2023).

The bidirectional long short-term memory (BiLSTM) paradigm is widely used in ER. Ye et al. (2023) proposed a text ER method based on an improved BiLSTM model and support vector machine (SVM)-Naive Bayes classification to get around the high subjectivity and poor recognition accuracy in text ER. Findings showed that this approach outperformed traditional modal recognition techniques in terms of accuracy (Ye et al., 2023). Mishra et al. (2024) proposed an integrated framework based on deep CNNs and BiLSTM to address the issues of low efficiency and lack of generalisation capability in traditional SER methods. Findings showed that a classification accuracy of up to 96.36% was attained by the hybrid deep neural network and BiLSTM model (Mishra et al., 2024). To address contextual understanding and category imbalance in tweet

sentiment detection, Kanam et al. (2025) proposed a novel hybrid deep learning architecture that combines synthetic minority class oversampling with gated recurrent units-BiLSTM. Results indicated that the gated recurrent units-BiLSTM model achieved a maximum accuracy of 89% (Kanam et al., 2025). To solve problems like inadequate feature extraction dimensions in EEG-based ER, Li et al. (2023) proposed a hybrid model that combines CNNs, self-attention mechanisms (AM), and BiLSTMs. Important information was highlighted by the self-AM, spatio-temporal properties were extracted by the CNN, and temporal relationships were preserved by the BiLSTM. The model outperformed other comparator models in terms of recognition performance, according to the results (Li et al., 2023). Murugaiyan et al. (2023) proposed a hybrid model to address the challenge of machines struggling to process human emotions and predict context-specific sentiments. Results indicated that this hybrid model outperformed other CNN variants (Murugaiyan et al., 2023).

In summary, existing SER methods demonstrate significant performance advantages across multiple publicly available standard speech emotion datasets. The majority of studies, however, do not take into consideration interference factors that exist in actual classroom environments, such as background noise and speech overlap brought on by multiple speakers speaking at once. The direct use of current models in real-world teaching situations is hampered by this limitation. Furthermore, a lot of ER techniques only use one feature parameter, ignoring the effect of several parameters on recognition performance. The study suggests an ER technique that combines multi-feature fusion (MFF) and enhanced segmentation-based clustering to overcome these difficulties. The objective is to develop an end-to-end framework that is especially made for real-world classroom settings and is able to identify the emotions of students. In order to tackle the problem of high-intensity noise interference in classrooms, this study innovates by putting forward an improved U-Net model that incorporates local loss supervision. By optimising weights across the network's various depth levels, this model effectively reduces noise in mixed speech signals. Additionally, to accurately determine the start and end times of speakers as well as their identities, the study uses a segmentation algorithm based on the Bayesian information criterion (BIC) in conjunction with K-means clustering. Lastly, the study integrates MFCC, formant parameters, and fundamental frequency to create a multidimensional feature space during the ER phase. A BiLSTM network with an AM is fed this feature space. The system achieves high-precision recognition of typical classroom emotions among students by dynamically capturing important information of emotional expressions through attention weights (AW).

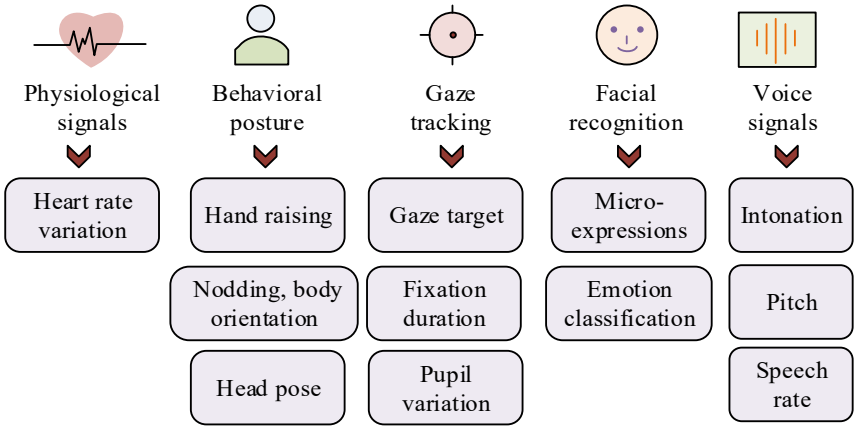
2 Methodology

The technical implementation pathways for classroom student ER are examined in this study, which also describes the main method designs for the three stages of speech denoising, segmentation and clustering, and ER. First, an enhanced U-Net model (I-U-Net-M) is suggested to improve speech signal quality in order to handle complex noise interference in classrooms. Second, a combination of BIC segmentation and K-means clustering is designed to precisely separate individual student speech segments. Lastly, a model that combines MFF with an improved BiLSTM is built to increase ER accuracy, creating an ER technology framework that is suited for actual classroom settings.

2.1 Classroom speech denoising based on an improved U-Net

The functional positioning of the classroom has changed from one-way knowledge transmission to a collaborative space for teachers and students to make meaning together as a result of the transformation of contemporary educational paradigms. Teachers must change from being traditional knowledge authorities to facilitators and learning process guides in order to make this shift. Teachers' primary duty is now to encourage the social construction of knowledge by encouraging students' intrinsic motivation and higher-order thinking, rather than just teaching them predetermined information. As a result, there has been a significant change in the ontological status of students in educational activities. The nature of knowledge changes during this process from static information that needs to be remembered and replicated to a dynamic medium that stimulates higher-order thinking and motivates problem-solving. The most obvious outward signs of the learning process in this dynamic teaching interaction are the emotional states and classroom engagement of the students. Positive classroom emotional experiences indicate that learners' cognitive systems have been fully activated, facilitating efficient learning. Conversely, negative emotional states signify excessive cognitive load, obstructing effective learning pathways. Therefore, the precise identification of these emotional states is a critical prerequisite for achieving personalised teaching and adaptive guidance (Akila et al., 2024; Anthony and Patil, 2023). Currently, intelligent classroom evaluation methods primarily rely on technical approaches based on computer vision and physiological sensing. The main characteristics of each approach are illustrated in Figure 1.

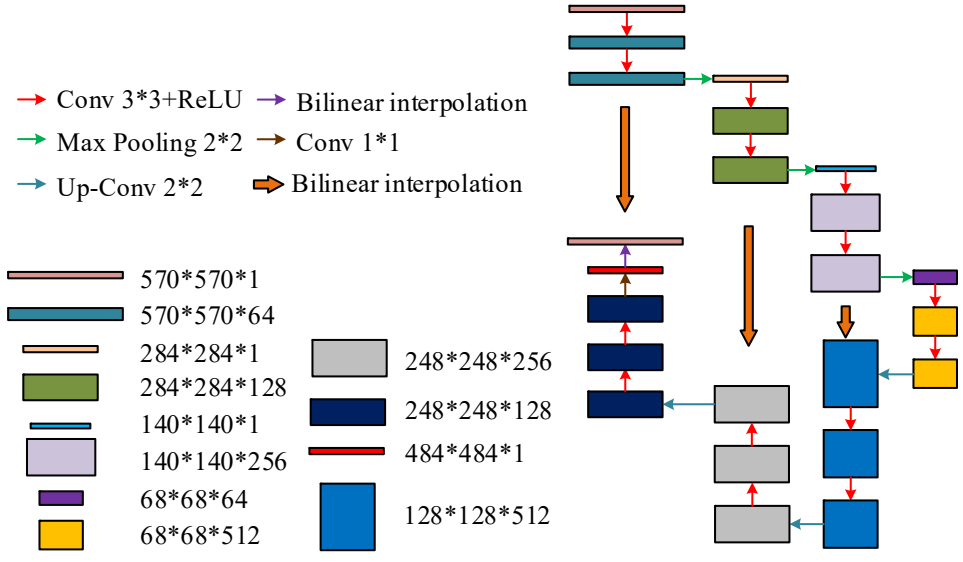
Figure 1 Characteristics of various classroom evaluation technology paths (see online version for colours)



In Figure 1, each technical approach to classroom evaluation has its own advantages and disadvantages. Methods based on physiological signals such as heart rate variability offer greater objectivity and lower costs, but their accuracy can be compromised by individual differences and factors like fatigue. Methods based on behavioural posture effectively reflect attention levels, but data collection in real offline classrooms is prone to incompleteness due to obstructions. Methods based on visual tracking directly reveal the focus of information acquisition, but they require expensive, sophisticated equipment and

impose stringent environmental requirements. Facial recognition-based methods are most closely associated with cognitive states and offer rich information, but they also face challenges in ensuring image clarity. Voice signal-based methods demonstrate superior privacy protection and, without requiring visual contact, can accurately reflect the speaker's engagement and emotional state (Wagner et al., 2023; Bhosle and Musande, 2023). The study primarily employs this method for classroom ER. Due to significant noise in classroom environments, speech denoising is first required (Talaat et al., 2023; Yan et al., 2023). To achieve effective speech denoising, the research proposes an I-U-Net-M that innovatively integrates local loss supervision and a weight update mechanism. The model's architecture is shown in Figure 2. In order to ensure the reproducibility of the model structure, the input speech spectrogram size was set to (1, 256, 256), and the output feature map sizes of the four stages of the encoder were (64, 128, 128), (128, 64, 64), (256, 32, 32), and (512, 16, 16), respectively. Correspondingly, the decoder gradually restores the dimensionality through upsampling, and finally outputs a denoised speech spectrogram that is consistent with the input size.

Figure 2 Structure diagram of I-U-Net-M (see online version for colours)



In Figure 2, the I-U-Net-M architecture incorporates operations such as convolution, deconvolution, max pooling, and bilinear interpolation. For stage loss computation, each deconvolution stage in the model is associated with a depth-wise loss function. This function ensures loss gradient backpropagation to accomplish weight updates by calculating the difference between the actual clean signal and the speech that has been reconstructed at this point. The drawbacks of depending only on global loss updates are addressed by the local weight update mechanism, which permits real-time adjustments during the encoder phase. It adapts to the demands of speech processing in noisy classrooms by controlling weights to enable precise feature map adjustments across decoder stages by utilising local losses from each upsampling stage. The deep supervision loss function is constructed in the study mainly using mean squared error (MSE) reconstruction error. The MSE reconstruction error is calculated as the loss value

at various model depth levels. The network is then guided to learn features at different depth levels by updating its weights using backpropagation gradients. The specific expression for MSE reconstruction error is shown in equation (1) (Xu et al., 2023).

$$L_{MSE}(y,\hat{y}) = \frac{1}{N}\sum_{a=1}^{N}(y_a - \hat{y}_a)^2 \tag{1}$$

In equation (1), L_{MSE} denotes the calculated MSE reconstruction error. a represents the sample point index. $\hat{y}$ indicates the predicted value. N signifies the total number of sample points. y is the actual value of the target signal. The core difference in improving U-Net in the study lies in the introduction of a multi-scale local loss supervision mechanism. Unlike standard U-Net, which only calculates losses at the final output layer, this model calculates local reconstruction errors at each upsampling stage of the decoder. Therefore, the overall optimisation objective of the model is no longer a single equation (1), but a weighted sum of losses in each stage. Within the encoder of the I-U-Net-M, the feature extraction convolutional module primarily employs convolutional layers for feature extraction, utilising the ReLU activation function to introduce nonlinear expressive capabilities. Following two rounds of convolutions, a deconvolution operation is performed. This operation primarily upsamples the feature maps generated by the first two convolutions, during which the initial loss of the encoding phase is computed. After completing two additional convolutions, loss calculation proceeds to further optimise the weights.

2.2 Classroom speech segmentation clustering based on BIC and K-means

After completing noise reduction processing on classroom audio, the next step involves segmenting and clustering the speech to identify speakers, time stamps, and spatial locations. This process comprises two components: segmentation and clustering. Considering the interactive nature of classroom audio, MFCC are selected as the feature parameters. Specifically, the frequency components of the denoised classroom audio signal are first converted into Mel frequencies, followed by calculating the cepstrum of these Mel frequencies to derive the MFCC. The expression for Mel frequencies is displayed in equation (2).

$$Mel(F) = 2595 \cdot \log_{10}\left(1 + \frac{F}{700}\right) \tag{2}$$

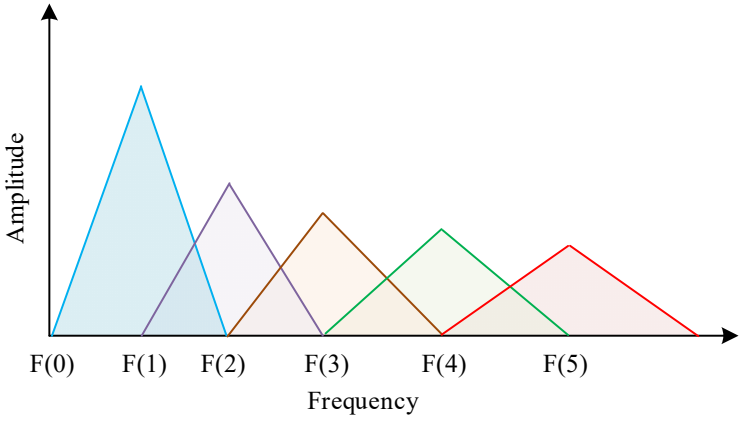
In equation (2), $Mel(F)$ is the Mel frequency. F is the actual frequency. Subsequently, the speech energy distribution is statistically analysed and subjected to Mel filter (MF). The MF bank consists of multiple triangular filters. Its schematic diagram is shown in Figure 3.

In Figure 3, the horizontal and vertical axis represents frequency and amplitude. Each triangular filter has its own centre frequency starting from the initial frequency $F(0)$. The filter amplitude first increases with frequency to reach a peak, then decreases, thereby achieving the filtering and extraction of speech energy across different frequency bands. In order to achieve the best feature extraction capability for the filter bank shown in Figure 3 in practical applications, a study was conducted to set the filter bank to include 40 triangular filters. The filter is nonlinearly distributed in the frequency range of

0 Hz–8,000 Hz, with dense distribution in the low-frequency region to capture fundamental details and sparse distribution in the high-frequency region, thus simulating the auditory perception characteristics of the human ear. The weighting coefficient $W_i(j)$ expression for the MF bank is displayed in equation (3).

$$W_i(j)=\begin{cases}\dfrac{j-F_c(i-1)}{F_c(i)-F_c(i-1)}, & F_c(i-1)\le j\le F_c(i)\\ 0, & j<F_c(i-1) \text{ or } j>F_c(i+1)\\ \dfrac{F_c(i+1)-j}{F_c(i+1)-F_c(i)}, & F_c(i)\le j\le F_c(i+1)\end{cases} \tag{3}$$

Figure 3 Schematic diagram of MF bank (see online version for colours)



In equation (3), i means the index of the filter. j means the index of the frequency point. $F_c(i)$ means the centre frequency of the i^{th} triangular filter. Next, the logarithmic energy $E(i)$ within each filter channel is calculated, as displayed in equation (4).

$$E(i)=\ln\left(\sum_{j=0}^{J-1}|S(j)|^2\,W_i(j)\right) \tag{4}$$

In equation (4), $S(j)$ represents the power spectrum of the original signal. J denotes the total number of frequency points in the spectrum. To reduce the correlation between features across dimensions, the logarithmic energy spectrum $E(i)$ is subsequently subjected to a discrete cosine transform, yielding the MFCCs. The calculation of the p^{th} cepstral coefficient M_p is displayed in equation (5).

$$M_p=\sum_{i=1}^{I}E(i)\cos\left(\frac{\pi p(i-0.5)}{I}\right) \tag{5}$$

In equation (5), I denotes the total quantity of triangular filters. Considering that speech is a continuously varying signal, the differential method is employed to capture its dynamic information in the time domain by obtaining the rate of change of MFCCs over time. The calculation of its first-order difference coefficient is shown in Equation (6).

$$\Delta_t(p) = \frac{\sum_{k=-L}^{L} k \cdot M_{t+k}(p)}{\sum_{k=-L}^{L} k^2} \tag{6}$$

In equation (6), $\Delta_t(p)$ means the dynamic change of the p^{th} MFCC feature at time t. $M_{t+k}(p)$ denotes the value of the p^{th} MFCC static feature at frame k relative to the current frame. L is a constant representing the range of context frames referenced during difference calculation, typically set to 2. Ultimately, the static MFCC features are combined with the calculated first-order and second-order difference features to form a sequence of feature vectors describing the speech signal. Considering the actual noise environment in classrooms, this study integrates BIC to achieve speaker segmentation. The algorithm primarily uses MFCC as the speech segmentation feature, treating the speaker segmentation task as a model selection process (Al-Dujaili and Ebrahimi-Moghadam, 2023). The calculation of BIC is shown in equation (7).

$$C_{BS} = M \ln(D) - 2\ln(\lambda_{\max}) \tag{7}$$

In equation (7), M denotes the total quantity of parameters in the current model. D represents the total scale of data points. $\lambda_{\max}$ indicates the maximum likelihood function obtainable under the current model. In acoustic modelling, the feature vector of a speech signal is displayed in equation (8).

$$\boldsymbol{v} \sim \mathcal{N}(\boldsymbol{m}, \boldsymbol{C}) \tag{8}$$

In equation (8), $\boldsymbol{v}$ is an d-dimensional eigenvector. $\mathcal{N}$ denotes a normal distribution. $\boldsymbol{m}$ and $\boldsymbol{C}$ displays the mean vector and covariance matrix of this distribution, respectively. The core of detecting speaker change points τ lies in testing two opposing hypotheses at a potential switching moment. If no switching exists, the entire speech sequence $\{\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_T\}$ is described by a single Gaussian distribution $\mathcal{N}(\boldsymbol{m}, \boldsymbol{C})$. If switching occurs, the speech sequence is split at time τ. The preceding portion $\{\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_\tau\}$ follows distribution $\mathcal{N}(\boldsymbol{m}_1, \boldsymbol{C}_1)$, while the subsequent portion $\{\boldsymbol{v}_{\tau+1}, \boldsymbol{v}_{\tau+2}, \ldots, \boldsymbol{v}_T\}$ follows another distribution $\mathcal{N}(\boldsymbol{m}_2, \boldsymbol{C}_2)$. The log-likelihood difference between the two is denoted as $\Delta\mathcal{L}(\tau)$, as shown in equation (9).

$$\Delta\mathcal{L}(\tau) = T\ln|C| - T_1\ln|C_1| - T_2\ln|C_2| \tag{9}$$

In equation (9), C, C_1, and C_2 represent the covariance matrices of the entire data segment, the pre-switch segment, and the post-switch segment, respectively. T, T_1, and T_2 denote the lengths of the corresponding data segments. The computed $\Delta\mathcal{L}(\tau)$ value is combined with the complexity penalty term and substituted into the BIC criterion to make the final determination on whether a speaker switch occurred. The BIC-based segmentation process is illustrated in Figure 4.

In Figure 4, the specific process of speech segmentation involves first initialising a small detection window and using the BIC algorithm to determine whether a segmentation point exists. If no segmentation point is found, the window's starting position remains unchanged, but its size gradually increases. When a segmentation point is detected, the window moves to a position after that point and resets to its initial smaller

size. This process repeats until the entire audio segment is analysed. Among them, the penalty factor in the BIC criterion is set to 1.2. During the speech clustering phase, the study primarily employs the K-means algorithm, with the specific workflow illustrated in Figure 5.

Figure 4 BIC based segmentation process (see online version for colours)

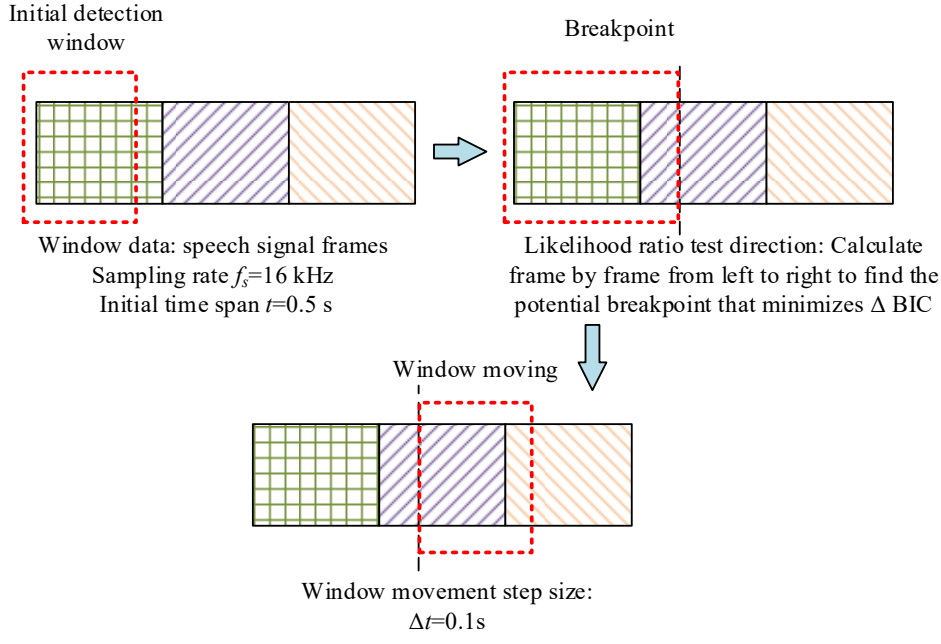
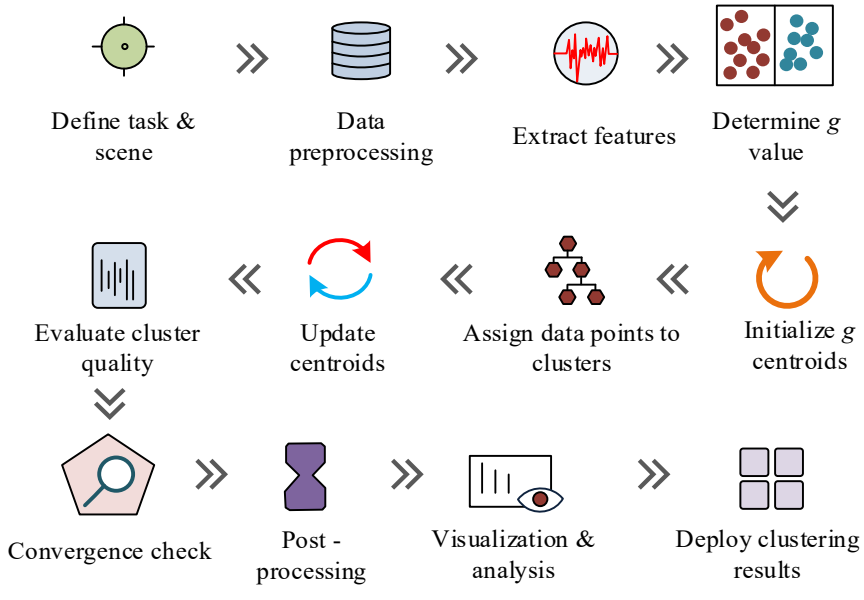


Figure 5 Clustering process based on K-means algorithm (see online version for colours)



In Figure 5, the speech clustering process first determines the quantity of clusters g. A quantity of data points equal to g are selected from the speech database as initial centroids. Subsequently, using the MFCC parameter matrices of each centroid as features, the speech data is clustered into g categories. If the termination condition is met, the algorithm stops. If not satisfied, each cluster re-selects its centroids and iteratively performs clustering operations. Through continuous updating of centroids and repeated clustering, the speech data is accurately categorised based on MFCC features. In terms of the specific parameter settings of the algorithm, in order to balance the granularity and computational efficiency of segmentation, the initial detection window size of the BIC algorithm is set to 1.5 s, and the window growth step is set to 1.5 s. For the determination of the number of clusters in K-means clustering, the elbow rule was used in the study. By calculating the sum of squared errors (SSE) within clusters with different numbers of clusters, the inflection point where the rate of SSE decrease significantly slows down is selected as the optimal number of clusters.

2.3 ER based on MFF and an improved BiLSTM

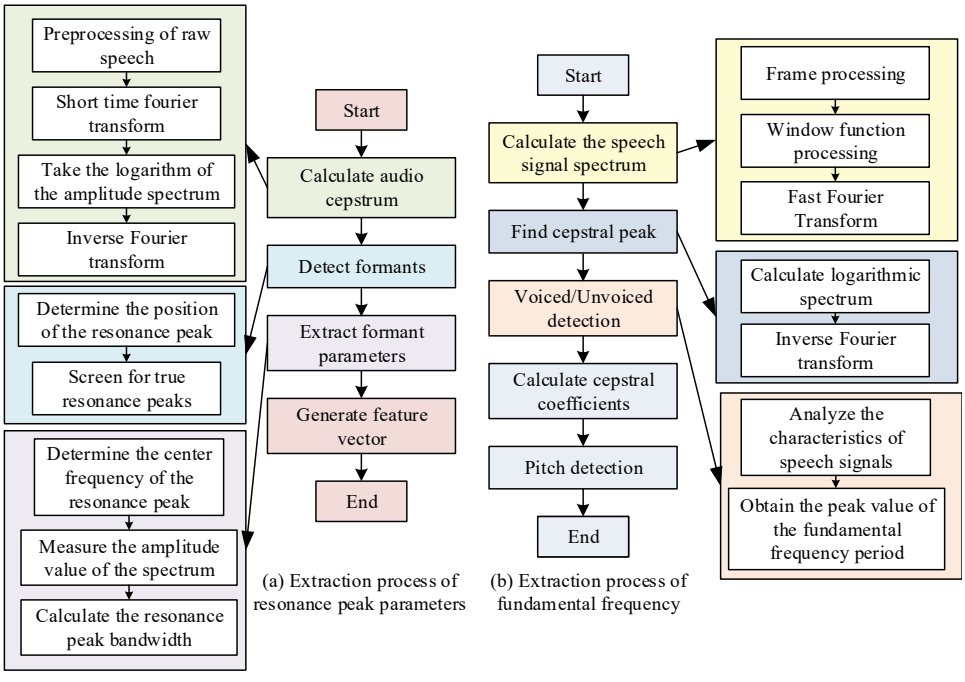
After completing the segmentation and clustering of classroom speech, the process proceeds to the ER stage. The selection of feature parameters in ER determines the quality of recognition performance. To address the problem of poor robustness in ER using single features, this study employs a fusion of MFCC, formant parameters, and fundamental frequency features, combined with speech segments obtained from classroom speech segmentation and clustering, to recognise the emotions of speaking students. In classroom speech, MFCC effectively captures emotional cues implied by the speech rhythm and intensity variations of the speaking student. Formant parameters reflect vocal tract resonance characteristics under different emotional states. Fundamental frequency enables emotional state assessment based on the speaking student's pitch. The MFCC extraction method is described above. Both formant parameters and fundamental frequency are extracted using the cepstral method. The specific extraction process is illustrated in Figure 6.

In Figure 6, the formant parameter extraction process begins with preprocessing the raw speech signal, including short-time Fourier transform (STFT), logarithmic amplitude spectrum extraction, and inverse Fourier transform (IFT). Subsequently, formant locations are identified, genuine formants are filtered, centre frequencies are determined, amplitudes are measured, and formant bandwidths are calculated. Fundamental frequency extraction begins by computing the speech signal spectrum, locating spectral peaks, performing voiced/unvoiced detection, calculating cepstral coefficients, conducting pitch detection, and concluding the process. The formant parameters and fundamental frequencies corresponding to different emotion types are shown in Table 1.

Table 1 Former and fundamental frequency corresponding to different emotional types

Emotional type	*Formant*		*Pitch frequency (Hz)*	
	Mean value of the first former	*Former variance*	*Mean fundamental frequency*	*Fundamental frequency variance*
Calm	17.21	5.47	147.96	123.00
Excitement	32.63	25.67	284.69	3480.00
Low	9.84	18.46	174.63	215.00

Figure 6 Extraction process of former and fundamental frequency (see online version for colours)



The study then performs feature fusion of MFCCs, formant parameters, and fundamental frequency for classroom student ER. By integrating three types of parameters, a richer emotional feature space can be constructed through complementary dimensions such as auditory simulation, pitch variation, and vocal tract resonance. Considering that these three features belong to distinct feature spaces, direct fusion may lead to conflicts. Therefore, the study employs a feature transformation approach, with the specific workflow being: first extracting feature vectors. Considering the significant differences in physical dimensions and numerical ranges among MFCC, former, and fundamental frequency, direct concatenation may result in features with larger values masking features with smaller values, leading to difficulties in model convergence. Therefore, before performing feature concatenation, the study adopts the maximum minimum normalisation method to map the three types of feature vectors to the [0, 1] interval respectively, in order to eliminate dimensional conflicts. Next, the three feature vectors are concatenated directly to achieve preliminary feature merging. Subsequently, to execute classroom SER activities, the concatenated vectors are mapped into a new feature space using principal component analysis, a dimensionality reduction technique. For ER in classroom students, this study proposes a BiLSTM network model incorporating an AM. In traditional BiLSTM gate structures, the weight matrix and bias matrix are fixed, resulting in equal learning weights assigned to information at each time step within a student's emotional feature sequence. However, in reality, emotional expressions among students in the classroom exhibit distinct dynamics. The significance of emotional information conveyed through facial expressions, body language, or vocal intonation varies considerably across different moments. Therefore, the study introduces an AM into the BiLSTM, using the

output sequence of the BiLSTM as input to the attention module. It employs a neural network-based additive attention calculation method to train the weights corresponding to each element in the sequence. The AM's calculating method can be broken down into the three main steps listed below. First, additive alignment is used to compute alignment scores, which measure the similarity between the query and each key, which expressed $E(\boldsymbol{Q}, \boldsymbol{K})$, as shown in equation (10).

$$E(\boldsymbol{Q}, \boldsymbol{K}) = \boldsymbol{\omega}_v^{\top} \tanh\left(\boldsymbol{\omega}_1 \boldsymbol{Q} + \boldsymbol{\omega}_2 \boldsymbol{K}\right) \tag{10}$$

In equation (10), $\boldsymbol{\omega}_1$ and $\boldsymbol{\omega}_2$ are trainable weight matrices. $\boldsymbol{\Omega}_v$ is a trainable weight vector. $\boldsymbol{Q}$ and $\boldsymbol{K}$ represent the query and key vectors, respectively. The research mainly uses the hidden state output of BiLSTM network in the last time step as a Query to represent the global semantic summary of the entire speech. And use the output sequence of BiLSTM at all time steps as keys and values. By calculating the interaction score between the Query and each Key, dynamically assigning AW to different time step values, effectively capturing key temporal segments of emotional expression. Subsequently, the alignment scores are normalised using the Softmax function to obtain the final AWs, as shown in equation (11).

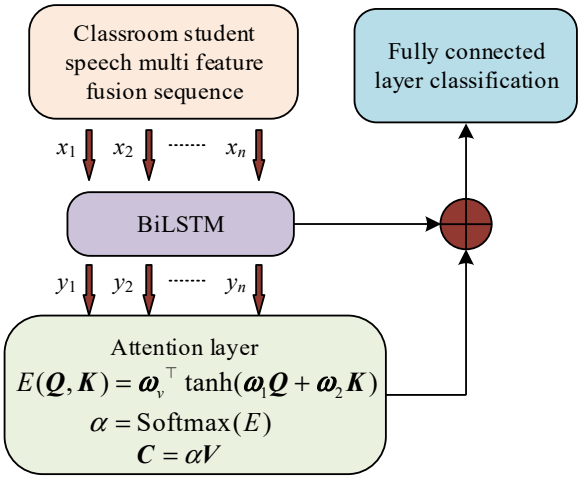
$$\alpha = \text{Softmax}(E) \tag{11}$$

In equation (11), α represents the final AWs. After processing through the Softmax function, all weight values fall within the interval [0, 1] and sum to 1. Finally, the corresponding values are weighted and summed using the obtained AWs α to generate the final context vector, as shown in equation (12).

$$\boldsymbol{C} = \alpha \boldsymbol{V} \tag{12}$$

In equation (12), $\boldsymbol{V}$ is the value vector, and $\boldsymbol{C}$ is the context vector. Figure 7 displays a schematic diagram of the attention-based BiLSTM.

Figure 7 Schematic diagram of BiLSTM based on AM (see online version for colours)



In Figure 7, for ER using the attention-based BiLSTM, the MFF sequence of classroom student speech is first input into the BiLSTM network. Subsequently, the processed

feature sequence is fed into the AM to compute the AWs for each output. The fused attention features are obtained through weighted summation. Finally, this feature is input into a fully connected layer to complete the ER task for classroom student speech. In addition, in order to enable the model to adapt to edge devices with limited computing resources and cope with the differences in acoustic environments of different classrooms, further model optimisation strategies were introduced in the method design. The improved BiLSTM constructed is defined as a lightweight student network that learns the output distribution of a complex teacher network using KL divergence loss function through knowledge distillation techniques. At the same time, in order to enhance environmental adaptability, the model integrates an incremental learning module that supports online fine-tuning of fully connected layer weights based on newly collected specific classroom audio data during the inference phase.

3 Results and analysis

The study conducts experimental validation and results analysis centred on the core components of the classroom student ER framework. Model performance is evaluated across three sequential stages: speech denoising, segmentation and clustering, and ER. First, addressing the challenge of complex classroom noise interference, the denoising effectiveness of the enhanced U-Net model is verified. Subsequently, based on the denoised speech signals, the accuracy and efficiency of the combined BIC segmentation algorithm and K-means clustering scheme are tested. Finally, for the core ER task, the study compares the recognition performance of MFF methods and an improved BiLSTM model.

3.1 Experimental analysis of speech denoising using an I-U-Net-M

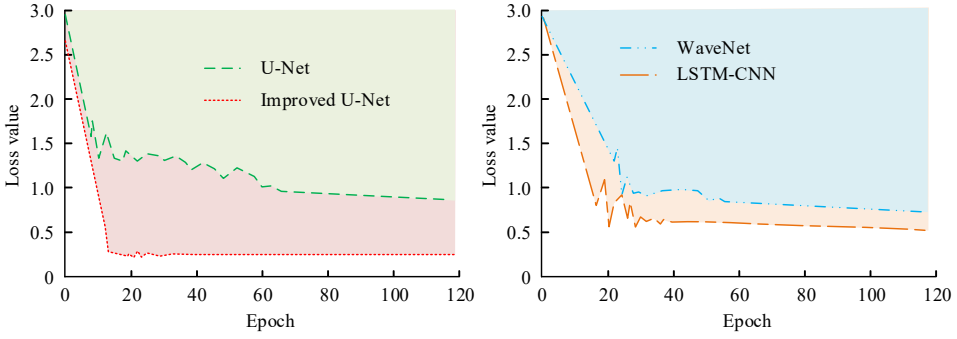
The total duration of the dataset used in the study is 80 hours. In terms of data distribution, pure speech was obtained from 60 student volunteers, including 30 boys and 30 girls, with an average age of 18.5 ± 1.8 years. Each participant contributed approximately 40 minutes of speech data, including three typical emotions of calmness, excitement, and depression, and the sample size of each emotion category remained balanced, ensuring a balanced distribution of the dataset in terms of gender, age, and individual duration. The classroom noise library includes six typical types of real noise, including environmental background noise, the sound of desks and chairs dragging, the sound of teachers lecturing, and the sound of multiple people chatting. To quantify the degree of noise interference and ensure experimental reproducibility, the study strictly mixed the above-mentioned types of noise with pure speech in three specific signal-to-noise ratio (SNR) intervals, including high noise mode (–5 dB–2 dB), medium noise mode (3 dB–8 dB), and low noise mode (10 dB–15 dB). This graded mixing strategy constructs a noisy mixed speech library covering from extreme interference to slight interference, effectively simulating the acoustic environment of a real classroom. Throughout the entire process of data collection and processing, the research strictly adheres to academic ethical standards and data privacy protection principles. Before the recording began, the students in the experiment signed an informed consent form. In order to ensure privacy compliance, all collected raw voice data undergoes strict anonymisation and desensitisation before being stored, retaining only emotion labels used

for acoustic analysis. The processed data is encrypted and stored on an offline server, authorised only for academic analysis by our research team. Table 2 displays the experimental setup.

Table 2 Experimental environment

Name	*Configuration*
Deep learning framework	Torch 1.7.0
Dependency library	CUDA 11.0
Language	Python 3.7.1
Operating system	Windows 11
Graphics processing unit	NVIDIA GeForce RTX 3060ti
System memory	16G
Central processing unit	AMD Ryzen 5600X

Figure 8 Training loss curves of different speech denoising models, (a) U-Net, improves U-net (b) WaveNet, LSTM-CNN (see online version for colours)

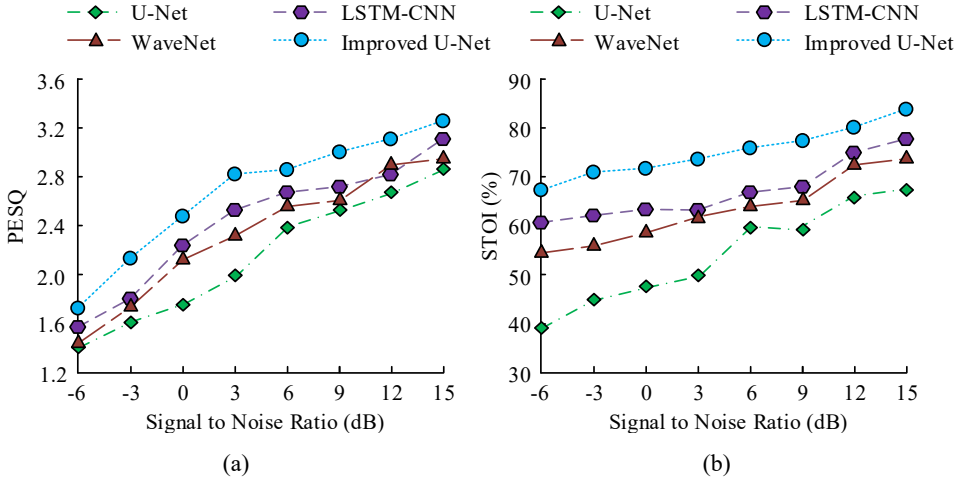


To validate the speech denoising performance of the I-U-Net-M, this study compares it with various noise reduction models, including the baseline U-Net, WaveNet, and long short-term memory-convolutional neural network (LSTM-CNN). Figure 8 displays each model's training loss curves. The I-U-Net-M exhibits the fastest convergence rate and the lowest loss value. It converges to 0.26 after approximately 16 iterations. In contrast, the baseline U-Net model requires 63 iterations to converge, with a loss value as high as 1.08. The final loss values for the WaveNet and LSTM-CNN models are 0.92 and 0.68, respectively, both higher than those of the modified U-Net model. Compared to the oscillation phenomenon observed in the early iterations of the basic U-Net, the improved model can quickly stabilise, indicating that the introduced local weight update mechanism effectively constrains the direction of gradient descent and prevents the network from getting trapped in local minima during optimisation.

The study then employs short-term objective comprehensibility (STOI) and perceived evaluation of speech quality (PESQ) to assess the noise reduction performance of each model. Higher PESQ scores indicate that the noise-reduced speech quality more closely resembles the original unadulterated speech, signifying superior noise reduction performance. Higher STOI scores represent greater intelligibility of the noise-reduced

speech. The study compares the PESQ and STOI scores of various models under different SNRs. In Figure 9(a), the PESQ scores of each model gradually increase with rising SNR. At an SNR of 9 dB, the I-U-Net-M achieves a PESQ of 3.01, representing a 0.48 improvement over the baseline U-Net. In contrast, WaveNet and LSTM-CNN yield PESQ scores of only 2.59 and 2.68, respectively. In Figure 9(b), the STOI of the I-U-Net-M consistently outperforms other models. At a SNR of 15 dB, the I-U-Net-M achieves an STOI of 84.21%, surpassing the LSTM-CNN model by 8.21 percentage points. The baseline U-Net model only achieves an STOI of 67.48%, confirming the superior speech denoising performance of the I-U-Net-M. The PESQ and STOI scores of all comparative models exhibit a positive correlation with SNR, showing a monotonic increase as SNR rises. This is because input signals with higher SNR preserve more original harmonic structures, thereby reducing the difficulty of neural network reconstruction of speech features. The PESQ and STOI indicators selected for the study are mainly used to evaluate the performance of the front-end speech enhancement module in waveform restoration and comprehensibility. Although ER tasks typically focus on acoustic feature sets, high-quality speech signals are a prerequisite for accurately extracting these fine-grained emotional features. Therefore, improving the excellent performance of U-Net on PESQ and STOI indirectly ensures the robustness of subsequent emotional features such as MFCC.

Figure 9 (a) PESQ and (b) STOI scores of various models under different SNRs (see online version for colours)



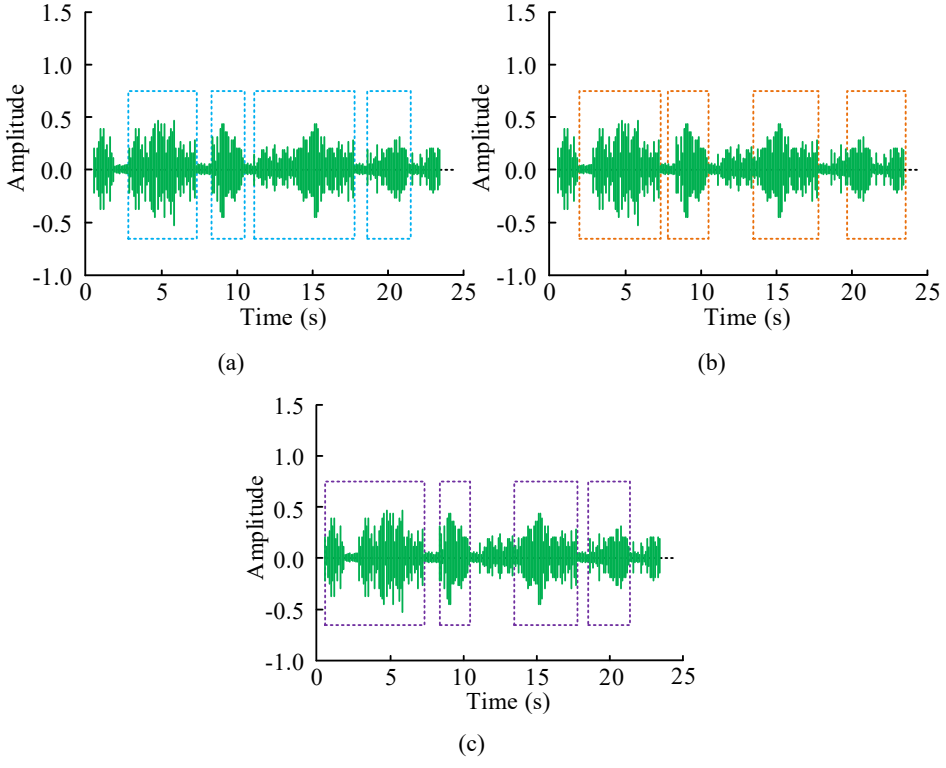
To further validate the effectiveness of the improved local loss supervision mechanism, a comparative experiment was conducted between the proposed model and the current mainstream cutting-edge speech enhancement models speech enhancement GAN (SEGAN) and Demucs. The performance comparison results of different models under the average SNR of the test set are shown in Table 3. According to Table 3, the PESQ of the improved U-Net reaches 3.01, which is 0.12 and 0.25 higher than Demucs and SEGAN, respectively. At the same time, the improvement of U-Net on STOI metrics is 2.79% higher than the best performing comparison model Demucs. In addition, in terms of inference time, the improved U-Net only takes 14.1ms, which is significantly better

than Demucs and SEGAN, and is more suitable for real-time processing needs in classroom scenarios.

Table 3 Comparison of denoising performance of mainstream models

Model	*PESQ*	*STOI (%)*	*Inference time (ms/frame)*
Basic U-Net	2.53	67.48	12.5
WaveNet	2.59	71.20	45.2
LSTM-CNN	2.68	76.00	18.4
SEGAN	2.76	79.15	32.6
Demucs	2.89	81.42	24.8
Improved U-Net	3.01	84.21	14.1

Figure 10 Segmentation maps of different speech segmentation models, (a) BIC model (b) Viterbi model (c) GMM-HMM model (see online version for colours)



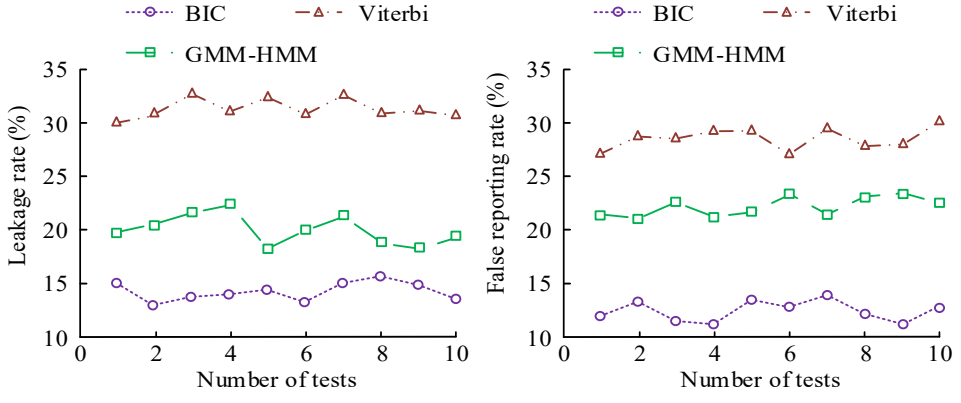
3.2 *Experimental analysis of speech segmentation and clustering in classroom settings*

To validate the effectiveness of the BIC segmentation model, the study employs the Viterbi model and the Gaussian mixture model-hidden Markov model (GMM-HMM) model for comparison. Figure 10 illustrates the speech segmentation results for different models. The BIC segmentation model accurately identifies the audio segments where

students speak during class discussions and detects transition points between speakers, enabling effective speech segmentation. Compared to the BIC model, the Viterbi model and GMM-HMM exhibit less accurate segmentation results during certain time intervals. These models may incorrectly classify non-speech segments as speech segments or misidentify speech segments as non-speech segments.

The study further employs the false negative rate and false positive rate as evaluation metrics for various speech segmentation models. In Figure 11(a), the BIC model exhibits a false negative rate of only 13.84%, whereas the Viterbi and GMM-HMM models show rates as high as 32.01% and 18.27%, respectively. In Figure 11(b), the false positive rate of the BIC model is only 12.52%, representing reductions of 16.20% and 9.28% compared to the Viterbi and GMM-HMM models, respectively. This demonstrates that the BIC model can accurately identify and segment classroom speech audio. In existing literature on speech segmentation for complex noisy environments, the average false positive and false negative rates of traditional methods are between 20% and 25%. The 12.52% false alarm rate obtained in the study has significantly decreased compared to these typical reference values, which strongly proves that the BIC algorithm has obvious advantages in suppressing over segmentation caused by classroom environment noise by introducing a model complexity penalty term.

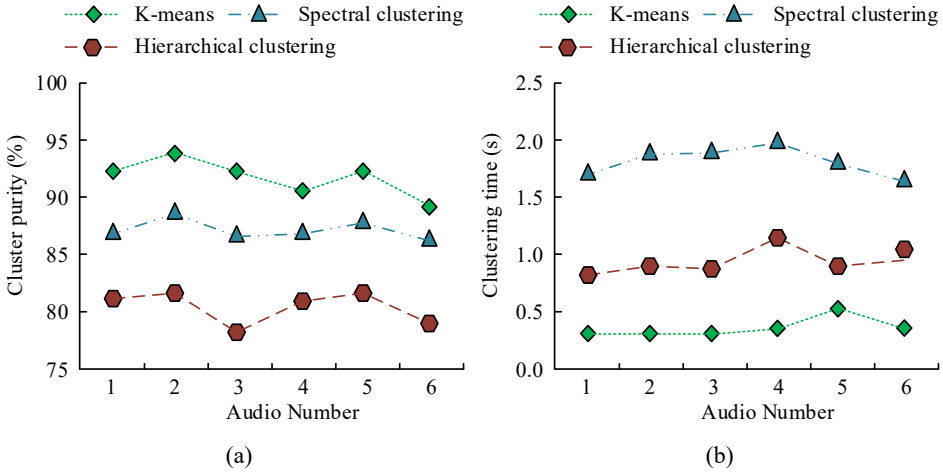
Figure 11 The omission rate and false alarm rate of each model, (a) leakage rates of different models (b) false reporting rate of different models (see online version for colours)



The study further verifies the superiority of the K-means clustering model. Using MFCC feature vectors extracted from segmented single-speech segments as clustering objects, spectral clustering and hierarchical clustering are selected as comparison models to evaluate two key metrics: clustering purity and clustering time. The experimental data consists of six typical multi-speaker classroom interaction audio segments from a noisy mixed speech database. The clustering purity and clustering time for each model are shown in Figure 12. In Figure 12(a), the clustering purity of the K-means clustering model reaches 90.85%, significantly higher than the 86.88% of spectral clustering and the 79.46% of hierarchical clustering. In Figure 12(b), the clustering time of the K-means clustering model is only 0.31 seconds, while spectral clustering takes 1.84 seconds and hierarchical clustering takes 0.93 seconds. Although spectral clustering performs well in purity indicators, its computational complexity results in a time consumption that is about

six times that of K-means, making it difficult to meet real-time requirements. However, hierarchical clustering has no advantage in both indicators. In contrast, the K-means clustering model always maintains the optimal balance between purity and speed, which is beneficial for real-time speech analysis in the classroom.

Figure 12 (a) Clustering purity and (b) clustering time of each model (see online version for colours)



3.3 Performance analysis of ER based on MFF and improved BiLSTM

To validate the effectiveness of the ER method combining MFF with an improved BiLSTM, the study first compares the recognition performance of different feature parameters for identifying student emotions during classroom participation. In Figure 13(a), the MFF model achieves the highest recognition accuracy for excited, depressed, and calm emotions, at 92.98%, 92.17%, and 89.64%, respectively. In contrast, the model using only MFCC results in recognition accuracy of only 78.64%, 68.72%, and 75.83%, respectively. The model utilising only gene frequency and formant parameters achieves maximum accuracies of merely 63.93% and 60.82%, respectively. In Figure 13(b), the MFF model achieves recall rates of 88.57%, 82.61%, and 87.24% for excited, depressed, and calm emotions, respectively, significantly outperforming other models. This demonstrates that the MFF model exhibits superior ER performance.

In order to further verify the necessity of multi feature fusion from the perspective of interpretability and quantify the actual contributions of each feature, the study introduced the SHapley Additive exPlans (SHAP) value for feature importance analysis. The average SHAP values and specific contributions of each feature are shown in Table 4. The analysis results show that in the fused feature space, the average SHAP value of MFCC features is the highest, reaching 0.45, indicating its dominant role in capturing speech spectral envelopes. Next is the fundamental frequency, with a SHAP of 0.32, which significantly contributes to distinguishing emotions with similar arousal levels such as calmness and depression. Formant has the smallest proportion, but provides crucial supplementary information in emotional variations of specific vowel pronunciations, confirming the complementarity of the three features in emotional representation.

Figure 13 Recognition performance of different feature parameters, (a) identification accuracy rate of different methods (b) recognition recall rate of different methods (see online version for colours)

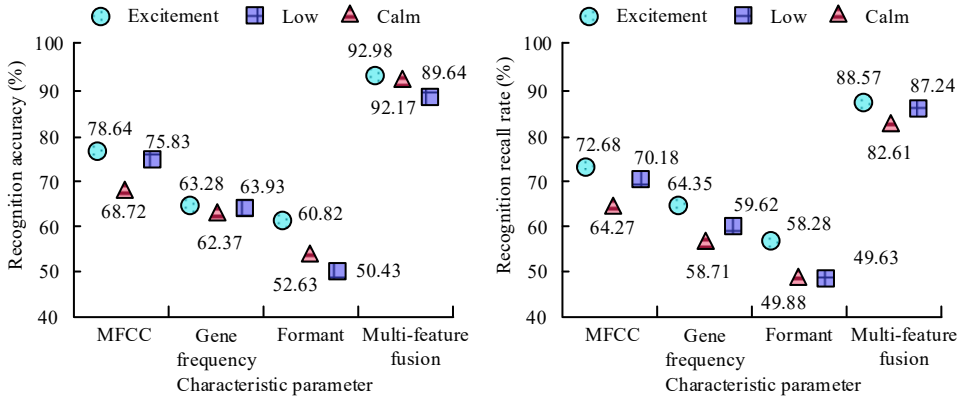
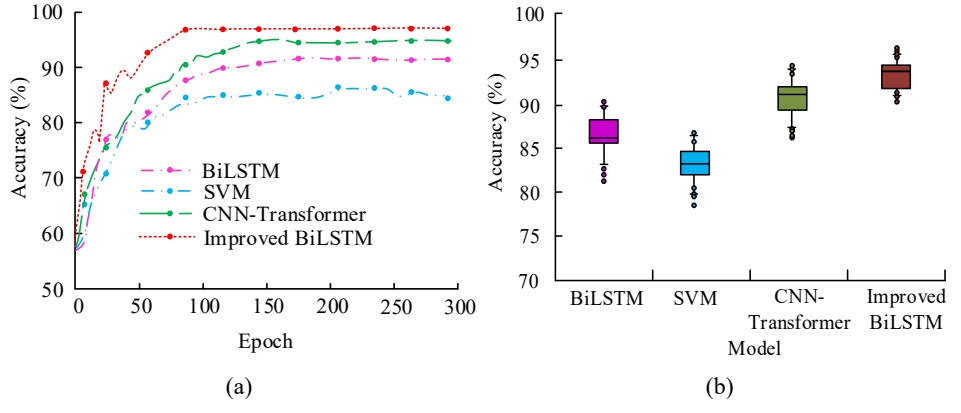


Table 4 SHAP value analysis of feature importance

Feature type	*Average SHAP value*	*Contribution rank*	*Primary contribution description*
MFCC	0.45	1	Capturing speech spectral envelope and timbre
Pitch frequency	0.32	2	Differentiating arousal levels
Formant	0.23	3	Reflecting vocal tract changes in specific vowels

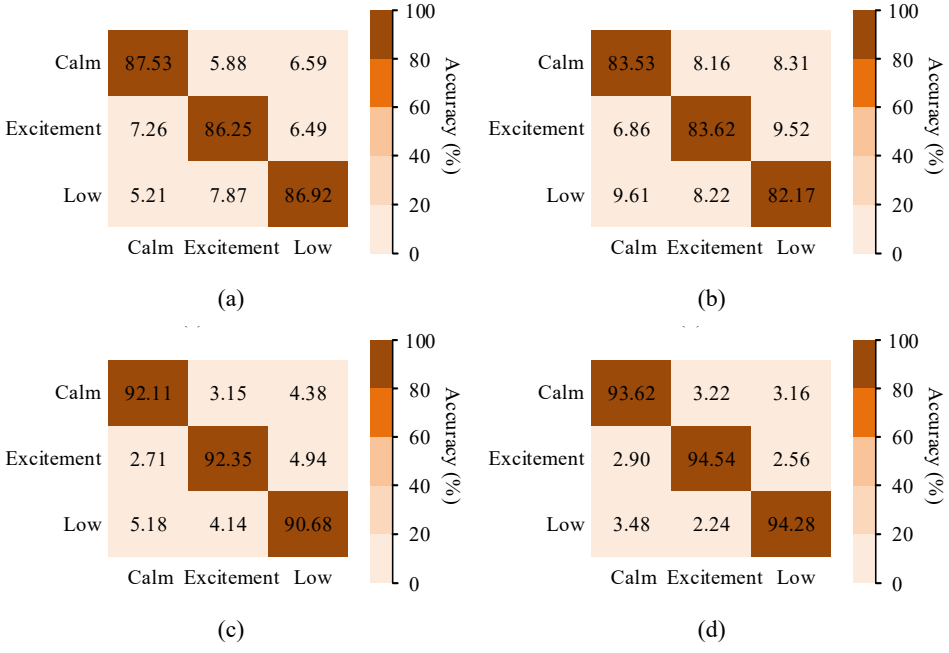
Figure 14 The accuracy of ER for each model, (a) training accuracy of different models (b) testing accuracy of different models (see online version for colours)



The study then validates the superiority of the improved BiLSTM model in ER by comparing it with traditional BiLSTM, SVM, and CNN-Transformer models. It first contrasts the convergence curves during training and the box plots during testing across all models. In Figure 14(a), the training accuracy curve of the improved BiLSTM model

converges to 97.24% after approximately 80 iterations. In contrast, the traditional BiLSTM model converges to 90.13% after 100 iterations. The SVM model converges to 84.63% after around 100 iterations. The CNN-Transformer model converges to 94.25% after approximately 150 iterations. In Figure 14(b), the improved BiLSTM model achieves a test accuracy of 93.62%, while the traditional BiLSTM, SVM, and CNN-Transformer models attain test accuracy of only 86.35%, 83.22%, and 92.13%, respectively. This demonstrates that the improved BiLSTM model outperforms the other comparison models in both ER accuracy and efficiency.

Figure 15 Confusion matrix of various models for ER of classroom students, (a) BiLSTM (b) SVM (c) CNN-transformer (d) improved BiLSTM (see online version for colours)



The study finally compares the confusion matrices of various models for recognising student emotions in the classroom. In Figure 15(a), the traditional BiLSTM model achieved recognition accuracy rates of 87.53%, 86.25%, and 86.92% for calm, excitement, and low mood, respectively. In Figure 15(b), the SVM model achieves recognition accuracy rates of only 83.53%, 83.62%, and 82.17% for each emotion, respectively. In Figure 15(c), the CNN-Transformer model achieves recognition accuracy rates of 92.11%, 92.35%, and 90.68% for each emotion, respectively. In Figure 15(d), the improved BiLSTM model achieves recognition accuracy rates of 93.62%, 94.54%, and 94.28%, respectively, significantly outperforming the other comparison models. The main recognition error of the model is concentrated between calm and low emotions. Mainly attributed to the high similarity in acoustic characteristics between the two. Calmness and depression both belong to low arousal emotions, with their mean fundamental frequencies being relatively close, and their energy variances being significantly lower than those of the excited state. Unlike the high-frequency and high-energy characteristics exhibited by excitement, distinguishing calmness from depression

is difficult to achieve solely based on global statistics, and requires subtle rhythmic dynamic changes. This further confirms the necessity of introducing attention mechanisms in research, that is, to compensate for the lack of discrimination between two types of emotions based on fundamental acoustic features such as fundamental frequency by capturing key differential frames in long time series.

In order to further evaluate the comprehensive advantages of the model in noise robustness and real-time performance, comparative experiments were conducted with the industry-leading end-to-end self-supervised pre training models Wav2Vec 2.0 and HuBERT. The comparison results of ER accuracy and real-time rate (real time factor, RTF) of different models on the test set are shown in Table 5. According to Table 5, Wav2Vec 2.0 and HuBERT, with their pre training advantages on large-scale unsupervised data, have slightly higher recognition accuracy than the improved BiLSTM model proposed by the research institute. However, the computational complexity of such end-to-end large models is extremely high, with RTF reaching 0.085 and 0.092 respectively, resulting in significant inference latency. In contrast, the improved BiLSTM model proposed in the study has an accuracy rate of up to 93.62%, only 0.53% lower than HuBERT, and an RTF of only 0.012, with an inference speed nearly 8 times faster than HuBERT. This indicates that the improved BiLSTM model has significant advantages in computational efficiency while ensuring high accuracy.

Table 5 Comparison of accuracy and real-time performance

Model	*Test accuracy (%)*	*Real-time factor (RTF)*
SVM	83.22	0.004
Standard BiLSTM	86.35	0.009
CNN-Transformer	92.13	0.025
Wav2Vec 2.0	93.89	0.085
HuBERT	94.15	0.092
Improved BiLSTM	93.62	0.012

To further validate the generalisation ability of the model in different language and cultural backgrounds, cross library validation experiments were conducted on the publicly available standard sentiment datasets EMODB and IEMOCAP. The experimental results are shown in Table 6. According to Table 6, the weighted accuracy (WA) of the improved BiLSTM model reached 88.45% on EMODB and 64.12% on IEMOCAP. This indicates that the multi feature fusion and attention mechanism framework constructed by the research institute has good cross domain robustness, not limited to specific classroom speech data.

In response to the demand for fine-grained learning analysis in smart teaching, the study expanded the recognition experiment of two classroom specific cognitive states, confusion and focus, on a supplementary test set containing annotations. The experimental results are shown in Table 7. According to Table 7, the F1 value of the model for the state of concentration is as high as 91.13%, indicating that the acoustic features of students when fully focused can be effectively captured by the model. Although the recognition difficulty of confused states is slightly higher, its F1 value still reaches 88.36%, indicating that the improved BiLSTM model has the potential for complex cognitive state recognition.

Table 6 Generalisation performance test results on public datasets

Dataset	*Language*	*WA*
EMODB	German	88.45%
IEMOCAP	English	64.12%

Table 7 Recognition performance of classroom specific states

State category	*Precision (%)*	*Recall (%)*	*F1 score (%)*
Confusion	89.24	87.50	88.36
Concentration	91.45	90.82	91.13
Average	90.35	89.16	89.75

4 Discussion and interpretation

The introduction points out that background noise interference and multi person speech mixing are two major challenges that hinder the application of existing ER technologies in real classrooms. Firstly, in response to strong classroom noise interference, the improved U-Net model designed in Section 2.1, which integrates local loss supervision, played a key role. The experimental results in section 3.1 show that even under low SNR conditions, the STOI and PESQ metrics of this model are still superior to baseline models such as LSTM-CNN. It has been confirmed that by introducing local supervision in the deep layers of the network, the model can effectively separate non-stationary noise during the feature encoding stage, thereby restoring clear classroom speech signals. Secondly, the BIC segmentation and K-means clustering strategies proposed in Section 2.2 provide effective solutions to the problem of speech aliasing caused by multi person conversations. The high clustering purity of 90.85% and the low dropout rate of 13.84% in section 3.2 indicate that this combination method can accurately segment mixed classroom audio into independent single speaker segments. In addition, in Section 2.3, attention mechanism was introduced to address the limitation of standard BiLSTM in distinguishing the importance of temporal information. From the confusion matrix, it can be seen that the improved model performs better in distinguishing emotions with similar acoustic features such as calmness and depression. This indicates that the attention mechanism successfully assigns higher weights to speech frames containing key emotional cues, enabling the model to capture subtle dynamic changes that are easily overlooked by standard BiLSTM from long sequences.

The results demonstrate that during the speech denoising phase, the proposed I-U-Net-M outperformed the baseline model in both training loss convergence speed and final performance metrics. Objective evaluation indicators PESQ and STOI achieved high values of 3.01 and 84.21%, respectively. The core mechanism behind this outcome was the innovative local loss supervision and weight update strategy. Traditional denoising models typically rely on global loss functions for end-to-end weight optimisation, leading to insufficient detail reconstruction when handling localised, sudden strong noise (Anees, 2024). The proposed I-U-Net-M achieves refined, phased control over feature maps at different network levels by introducing depthwise supervision loss at each upsampling stage of the decoder.

During the classroom speech segmentation and clustering phase, the study validated the effectiveness of combining the BIC segmentation algorithm with K-means clustering. Experimental data revealed that the BIC model exhibited significantly lower false negative and false positive rates compared to the Viterbi and GMM-HMM models. By introducing a data-volume-dependent penalty term to balance model fit and complexity, BIC effectively avoided over-segmentation caused by noise or brief silences. This approach yielded greater accuracy than Viterbi and GMM-HMM models, which relied solely on acoustic feature transfers. In the clustering phase, the K-means algorithm achieved a clustering purity of 90.85% with a processing time of just 0.31 seconds. This indicated that speech features from different speakers exhibit excellent separability in the MFCC feature space, where a simple Euclidean distance metric proved sufficient. For classroom applications requiring real-time feedback, K-means' computational efficiency could be its decisive advantage over more complex algorithms like spectral clustering. A. Gupta and Purwar's (2024) research primarily assumed input consisting of single-person, segmented pure speech. In contrast, this study successfully overcame the challenge of automatically segmenting speech from different speakers and performing emotion clustering in noisy, real-world, mixed-audio classroom recordings containing multiple conversations. This was achieved through highly efficient and precise segmentation and clustering (Gupta and Purwar, 2024).

In the core ER performance validation, the proposed MFF method combined with an improved BiLSTM model achieved higher recognition accuracy for all emotion categories than any single feature. This was based on the complementary nature of the features. MFCC modelled the auditory perception characteristics of the human ear. The formant parameters reflected the resonant differences in the physical structure of the vocal tract, while the fundamental frequency directly correlated with variations in pitch. These three components captured emotion from three distinct dimensions: perception, physiology, and prosody. Together, they constructed a more comprehensive emotional representation space than any single feature alone, validating the view of Benzirar et al. (2025) on the importance of feature fusion. Secondly, the improved BiLSTM model incorporating an AM ultimately achieved a test accuracy of 93.62%. Its superiority stemmed from the AM's ability to dynamically focus on key information. Traditional BiLSTM treated all time steps equally when processing sequences, which did not align with the actual nature of emotional expression. The AM enabled the model to automatically focus on frames within the speech sequence that contributed most significantly to emotion classification by learning a weight distribution, thereby assigning them higher weights. This approach more effectively captured the dynamics and critical details of emotional expression, representing a significant improvement over the BiLSTM application by Feng et al. (2024).

At the practical application level, although this study achieved efficient clustering in 0.31 seconds on a general PC platform, further consideration of computing power limitations is needed when deploying embedded recording terminals for classrooms. The future engineering landing can reduce the computational complexity of improved U-Net and BiLSTM through model pruning and quantification technology to meet the real-time requirements of edge computing devices. In addition, direct input into the model may lead to emotional misjudgement in response to common nonverbal sudden disturbances in the classroom. To enhance the robustness of the system, it is recommended to add an abnormal audio filtering module between speech denoising and segmentation in the actual deployment framework. This module can be based on dual threshold detection of

short-term energy and zero crossing rate, pre removing high-energy non speech segments to avoid interference with subsequent BIC segmentation and ER.

5 Conclusions

To achieve effective ER for students in the classroom, this study proposed a classroom student ER model that integrates an improved segmentation clustering approach with a MFF ER algorithm. Results indicated that during the speech denoising phase, the training loss of the proposed I-U-Net-M rapidly converged to 0.26 after approximately 16 iterations, outperforming traditional U-Net and WaveNet models. In the speech segmentation clustering experiment, the BIC-based segmentation model achieved low false negative and false positive rates of 13.84% and 12.52%, respectively, significantly outperforming traditional methods such as Viterbi. The K-means clustering algorithm achieved a high clustering purity of 90.85% with a clustering time of only 0.31 seconds. This validated the high accuracy and efficiency of the combined approach in precisely identifying speakers and their start/end times. In the core ER stage, the MFF approach achieved an accuracy rate of 92.98% for identifying excitement, significantly surpassing the 78.64% accuracy of the single MFCC feature. The improved BiLSTM model equipped with an AM attained an overall accuracy of 93.62% on the test set, outperforming both the traditional BiLSTM model (86.35%) and the CNN-Transformer model (92.13%). In summary, by integrating an I-U-Net-M with local loss supervision, a segmentation clustering algorithm based on BIC and K-means, and a BiLSTM recognition network combining MFF and AMs, this study successfully constructs an end-to-end student ER framework capable of effectively addressing the challenges posed by complex acoustic environments in real classrooms. This framework demonstrates high accuracy and strong robustness across noise reduction, segmentation, and recognition stages, offering significant practical application value.

Declarations

The author declares no conflicts of interest.

References

Akila, D., Garg, H., Pal, S. and Jeyalaksshmi, S. (2024) ‘Research on recognition of students attention in offline classroom-based on deep learning’, *Educ. Inf. Technol*., August, Vol. 29, No. 6, pp.6865–6893, DOI: 10.1007/s10639-023-12089-6.

Aldhilan, D. and Rafiq, S. (2025) ‘Transforming early childhood education in Saudi Arabia: AI’s impact on emotional recognition and personalized learning’, *Int. J. Eval. Res. Educ*., August, Vol. 14, No. 4, pp.2473–2486, DOI: 10.11591/ijere.v14i4.32660.

Al-Dujaili, M.J. and Ebrahimi-Moghadam, A. (2023) ‘Speech emotion recognition: a comprehensive survey’, *Wireless Pers. Commun*., March, Vol. 129, No. 4, pp.2525–2561, DOI: 10.1007/s11277-023-10244-3.

Anees, M. (2024) ‘Speech coding techniques and challenges: a comprehensive literature survey’, *Multimed. Tools Appl*., September, Vol. 83, No. 10, pp.29859–29879, DOI: 10.1007/s11042-023-16665-3.

Anthony, A.A. and Patil, C.M. (2023) 'Speech emotion recognition systems: a comprehensive review on different methodologies', *Wireless Pers. Commun*., March, Vol. 130, No. 1, pp.515–525, DOI: 10.1007/s11277-023-10296-5.

Ashok Kumar, P.M., Maddala, J.B. and Martin Sagayam, K. (2023) 'Enhanced facial emotion recognition by optimal descriptor selection with neural network', *IETE J. Res*., March, Vol. 69, No. 5, pp.2595–2614, DOI: 10.1080/03772063.2021.1902868.

Benzirar, A., Hamidi, M. and Bouami, M.F. (2025) 'Conception of speech emotion recognition methods: a review', *Indones. J. Electr. Eng. Comput. Sci*., March, Vol. 37, No. 3, pp.1856–1864, DOI: 10.11591/ijeecs.v37.i3.

Bhosle, K. and Musande, V. (2023) 'Evaluation of deep learning CNN model for recognition of Devanagari digit', *Artif. Intell. Appl*., February, Vol. 1, No. 2, pp.114–118, DOI: 10.47852/bonviewAIA3202441.

Bie, M., Liu, Q., H. Xu, Q., Gao, Y. and Che. X. (2024) 'FEMFER: Feature enhancement for multi-faces expression recognition in classroom images', *Multimed. Tools Appl*., May, Vol. 83, No. 2, pp.6183–6203, DOI: 10.1007/s11042-023-15808-w.

Chen, W., Xing, X., Chen, P. and Xu, X. (2024) 'Vesper: a compact and effective pretrained model for speech emotion recognition', *IEEE Trans. Affect. Comput*., July–September, Vol. 15, No. 3, pp.1711–1724, DOI: 10.1109/TAFFC.2024.3369726.

Ezquerra, A., Agen, F., Toma, R.B. and Ezquerra-Romano, I. (2025) 'Using facial emotion recognition to research emotional phases in an inquiry-based science activity', *Res. Sci. Technol. Educ*., July, Vol. 43, No. 1, pp.62–85, DOI: 10.1080/02635143.2023.2232995.

Falahzadeh, M.R., Farokhi, F., Harimi, A. and Sabbaghi-Nadooshan, R. (2023) 'Deep convolutional neural network and gray wolf optimization algorithm for speech emotion recognition', *Circuits Syst. Signal Process*., August, Vol. 42, No. 1, pp.449–492, DOI: 10.1007/s00034-022-02130-3.

Feng, X., Angkawisittpan, N. and Yang, X. (2024) 'A CNN-BiLSTM algorithm for Weibo emotion classification with attention mechanism', *Math. Model. Eng*., April, Vol. 10, No. 2, pp.87–97, DOI: 10.21595/mme.2024.24076.

Gupta, A. and Purwar, A. (2024) 'Speech refinement using Bi-LSTM and improved spectral clustering in speaker diarization', *Multimed. Tools Appl*., December, Vol. 83, No. 18, pp.54433–54448, DOI: 10.1007/s11042-023-17017-x.

Kanam, M., Munawir, M. and Efrizoni, L. (2025) 'Improved performance of hybrid GRU-BiLSTM for detection emotion on twitter dataset', *J. Appl. Data Sci*., May, Vol. 6, No. 1, pp.354–365, DOI: 10.47738/jads.v6i1.459.

Li, X., Li, P. and Fang, Z. (2023) 'Research on EEG emotion recognition based on CNN+ BiLSTM+ self-attention model', *Optoelectron. Lett*., August, Vol. 19, No. 8, pp.506–512, DOI: 10.1007/s11801-023-2207-x.

Mishra, S., Bhatnagar, N. and Prakasam, P. (2024) 'Speech emotion recognition and classification using hybrid deep CNN and BiLSTM model', *Multimedia Tools Appl*., October, Vol. 83, No. 13, pp.37603–37620, DOI: 10.1007/s11042-023-16849-x.

Mishra, S.P., Warule, P. and Deb, S. (2024) 'Speech emotion recognition using MFCC-based entropy feature', *Signal, Image Video Process*., August, Vol. 18, No. 1, pp.153–161, DOI: 10.1007/s11760-023-02716-7.

Murugaiyan, S. and Uyyala, S.R. (2023) 'Aspect-based sentiment analysis of customer speech data using deep convolutional neural network and BiLSTM', *Cogn. Comput*., March, Vol. 15, No. 3, pp.914–931, DOI: 10.1007/s12559-023-10127-6.

Panda, S.K., Jena, A.K. and Panda, M.R. (2023) 'Speech emotion recognition using multimodal feature fusion with machine learning approach', *Multimedia Tools Appl*., April, Vol. 82, No. 27, pp.42763–42781, DOI: 10.1007/s11042-023-15275-3.

Patnaik, S. (2023) 'Speech emotion recognition by using complex MFCC and deep sequential model', *Multimedia Tools Appl*., September, Vol. 82, No. 8, pp.11897–11922, DOI: 10.1007/s11042-022-13725-y.

Subramanian, R. and Aruchamy, P. (2024) 'An effective speech emotion recognition model for multi-regional languages using threshold-based feature selection algorithm', *Circuits Syst. Signal Process*., December, Vol. 43, No. 4, pp.2477–2506, DOI: 10.1007/s00034-023-02571-4.

Talaat, T.F.M., El-Gendy, E.M., Saafan, M.M. and Gamel, S.A. (2023) 'Utilizing social media and machine learning for personality and emotion recognition using PERS', *Neural Comput. Appl*., December, Vol. 35, No. 33, pp.23927–23941, DOI: 10.1007/s00521-023-08962-7.

Tang, X., Gong, Y., Xiao, Y., Xiong, J. and Bao, L. (2025) 'Facial expression recognition for probing students' emotional engagement in science learning', *J. Sci. Educ. Technol*., August, Vol. 34, No. 1, pp.13–30, DOI: 10.1007/s10956-024-10143-7.

Wagner, J., Triantafyllopoulos, A., Wierstorf, H., Schmitt, M., Burkhardt, F., Eyben, F. and Schuller, B. (2023) 'Dawn of the transformer era in speech emotion recognition: closing the valence gap', *IEEE Trans. Pattern Anal. Mach. Intell*., September, Vol. 45, No. 9, pp.10745–10759, DOI: 10.1109/TPAMI.2023.3263585.

Xu, Y., Su, H., Ma, G. and Liu, X. (2023) 'A novel dual-modal emotion recognition algorithm with fusing hybrid features of audio signal and speech context', *Complex & Intell. Syst*., August, Vol. 9, No. 1, pp.951–963, DOI: 10.1007/s40747-022-00841-3.

Yan, X., Lin, Z., Lin, Z. and Vucetic, B. (2023) 'A novel exploitative and explorative GWO-SVM algorithm for smart emotion recognition', *IEEE Internet Things J*., 1 June, Vol. 10, No. 11, pp.9999–10011, DOI: 10.1109/JIOT.2023.3235356.

Ye, Z., Zuo, T., Chen, W., Li, Y. and Lu, Z. (2023) 'Textual emotion recognition method based on ALBERT-BiLSTM model and SVM-NB classification', *Soft Comput*., February, Vol. 27, No. 8, pp.5063–5075, DOI: 10.1007/s00500-023-07924-4.

Zhang, Q., Zhang, H., Zhou, K. and Zhang, L. (2023) 'Developing a physiological signal-based, mean threshold and decision-level fusion algorithm (PMD) for emotion recognition', *Tsinghua Sci. Technol*., August, Vol. 28, No. 4, pp.673–685, DOI: 10.26599/TST.2022.9010038.

Zhu, X., Huang, Y., Wang, X. and Wang, R. (2024) 'Emotion recognition based on brain-like multimodal hierarchical perception', *Multimedia Tools Appl*., December, Vol. 83, No. 18, pp.56039–56057, DOI: 10.1007/s11042-023-17347-w.