International Journal of Trend in Scientific Research and Development (IJTSRD)
Volume 10 Issue 5, Sep-Oct 2026 Available Online: www.ijtsrd.com e-ISSN: 2456 – 6470

Design and Implementation of a Classroom Participation Evaluation System Based on Multimodal Fusion

Wenrong Song, Rui Dai, Zaiyuan Dong

School of Computer Science and Artificial Intelligence, Beijing Wuzi University, Beijing, China

ABSTRACT

Classroom participation is often treated as a key indicator of teaching quality, yet traditional evaluation relies mainly on teachers' on-site subjective observation, which suffers from limited evaluation dimensions, untimely feedback, and a heavy workload for teachers. Existing studies mostly focus on a single modality and lack deep fusion. Accordingly, this paper designs and implements a classroom participation evaluation system that combines speech recognition with image recognition. The system adopts a dual-path multimodal perception architecture. On the image path, SSD is first used for face detection, then MobileFaceNet outputs eight emotion categories, and MediaPipe Face Landmarker is used to obtain 478 three-dimensional facial landmarks for lightweight real-time tracking. On the speech path, librosa is used to extract multi-dimensional acoustic features such as fundamental frequency, RMS energy, and zero-crossing rate to classify emotional tendencies. At the decision level, a confidence-weighted fusion mechanism is introduced to integrate indicators including head-up rate, number of speaking turns, and expression positivity, producing multi-dimensional participation scores. The system is developed with open-source frameworks such as PyTorch, OpenCV, and uni-app, and all core computations are performed locally. Test results show that the facial expression recognition accuracy is 70.0 percent, the speech emotion classification consistency reaches 86.1 percent, the System Usability Scale score is 74.5, and memory usage remains stable within 120 MB during 60 minutes of continuous operation. This work provides a deployable technical solution for the automated evaluation of classroom participation in smart education environments.

How to cite this paper: Wenrong Song | Rui Dai | Zaiyuan Dong "Design and Implementation of a Classroom Participation Evaluation System Based on Multimodal Fusion" Published in International Journal of Trend in Scientific Research and Development (ijtsrd), ISSN: 2456-6470, Volume-10 | Issue-5, October 2026, pp.1878-1886, URL: www.ijtsrd.com/papers/ijtsrd150281.pdf

IJTSRD150281

Copyright © 2026 by author (s) and International Journal of Trend in Scientific Research and Development Journal. This is an Open Access article distributed under the terms of the Creative Commons Attribution License (CC BY 4.0) (http://creativecommons.org/licenses/by/4.0)

KEYWORDS: *multimodal fusion; classroom participation; speech recognition; image recognition; decision-level fusion.*

I. INTRODUCTION

Classroom participation reflects the degree of students' involvement in the teaching process and serves as an important basis for teachers to adjust their teaching strategies. In a traditional classroom, however, one teacher often faces dozens of students at the same time, making it difficult to continuously and carefully observe and record each student's participation. Such an evaluation method is not only inefficient but also easily influenced by the evaluator's subjective factors, so the objectivity of the evaluation results is hard to guarantee. In recent years, educational informatization has advanced rapidly, and using technical means to automatically analyze students' classroom behavior and provide data support for teaching reflection has become an important research direction in the field of smart education.

Advances in artificial intelligence technology provide strong support for classroom behavior analysis. In computer vision, techniques such as face detection, expression recognition, and posture estimation are gradually maturing; in speech processing, capabilities such as speech-to-text, voiceprint recognition, and speech emotion analysis are also improving. Yet many existing studies still concentrate on the application of a single modality. For example, some systems rely only on facial expressions to judge students' concentration, but in real classrooms occlusion of faces, changes in illumination, and students lowering their heads are common, so the reliability of a single modality is hard to guarantee. Other systems rely only on voice sign-in or behavior recognition, and their adaptability to scenarios is equally limited. Integrating speech and image

information at the decision level, and exploiting the complementarity of the two modalities to overcome the limitations of a single modality, is an effective way to improve the accuracy and robustness of classroom participation evaluation.

Based on the above considerations, this paper designs a multimodal participation evaluation system for classroom scenarios. The system integrates speech recognition and image recognition technologies and comprehensively analyzes students' participation from multiple dimensions such as attendance status, classroom speech, facial expressions, and speech emotion. It adopts a decision-level fusion strategy to generate participation scores and presents the analysis results to teachers through a visualization interface. The following sections introduce the related work, the overall system design, the key technologies, the experimental results, and the conclusions in turn.

II. RELATED WORK

A. *Image-Recognition-Based Classroom Behavior Analysis*

Research on image recognition in classroom scenarios is already fairly rich. In the direction of classroom attendance, some studies have built face recognition systems using the deep learning method of the Dlib library, improving both recognition speed and accuracy; other work has developed a lightweight classroom face sign-in system based on Python and OpenCV to solve the problems of proxy sign-in and time-consuming attendance. The common idea of these two approaches is to replace traditional name calling with face recognition, but in real classroom environments, side faces, occlusion, and illumination changes still cause obvious fluctuations in the recognition rate [4].

In the direction of classroom behavior recognition, a student behavior analysis system based on YOLOv5 realizes behavior detection through self-trained weights [5]. Subsequent research introduced a lightweight coordinate attention mechanism to improve the YOLOv8s backbone network, enhancing feature-learning ability [13]. The DMS-YOLOv8 algorithm is specifically designed to address the difficulty of detecting small targets in the back rows of smart classrooms, combining the coordinate attention mechanism with depth-wise convolution to design dynamic channel attention convolution [6]. However, such methods mostly output behavior category labels and do not directly map to quantifiable participation indicators. From detecting a lowered head to judging low participation, an intermediate scoring logic is still needed.

In the direction of classroom concentration detection, existing studies have attempted to combine facial expressions with behavioral features: some studies have addressed the limitations of facial expression recognition in real teaching environments and proposed a detection system that combines facial expressions with behavioral features, computing concentration scores through a fuzzy comprehensive evaluation algorithm [8]; other studies have constructed a concentration recognition model by combining classroom expressions and head posture, and further analyzed the influence of teachers' teaching gestures on learners' concentration [7]. Related work on deep-learning-based recognition of student classroom participation provides a further reference [9]. These works reflect that this field is shifting from single-dimensional analysis to multi-dimensional comprehensive evaluation.

B. *Speech-Recognition-Based Classroom Interaction Analysis*

Speech recognition has also made progress in classroom scenarios. In the direction of voiceprint recognition and classroom attendance, a Raspberry Pi-based scheme aims at the problems that traditional name calling is time-consuming and easily forged, obtains speech signals through a microphone to complete voiceprint comparison, and provides a non-contact solution for classroom attendance [10]. In the direction of speech emotion analysis, affective computing can mine the inner state of speakers from speech signals [11]. After extracting acoustic features such as fundamental frequency, energy, zero-crossing rate, and speaking rate, the emotional tendency can be judged automatically [12]. The mean and standard deviation of the fundamental frequency are significantly related to emotional states; the dynamic change of RMS energy converted to dB reflects the intensity and fluctuation of speech; and the zero-crossing rate can distinguish the degree of tension in speech. Studies on fusing text and acoustic features for classroom conversation emotion recognition further extend the acoustic-analysis direction [14]. It should be noted that simultaneous speech of multiple people and background noise in classroom environments will significantly affect the stability of these features, so speech enhancement and speaker separation are still needed in actual deployment.

C. *Multimodal Fusion Techniques*

The core idea of multimodal fusion is to use the complementarity of different data sources to overcome the limitations of a single modality. The decision-level fusion face-voice recognition method proposed by Li and Zhou has methodological reference value: the face image adopts a combination of principal component analysis and linear discriminant analysis, the speech information adopts

Mel-frequency cepstral coefficients and a Gaussian mixture model, and the outputs of the two subsystems are used as inputs to a support vector machine for decision-level fusion, effectively improving the overall recognition rate [1]. The inspiration of this approach is that fusion occurs at the decision level rather than the feature level, avoiding the interference caused by inconsistent data scales of different modalities.

In the field of classroom analysis, the application of multimodal fusion is receiving increasing attention. Based on the theory of multimodal learning analysis, Wu et al. analyzed learning engagement from three dimensions of behavior, cognition, and emotion, and constructed a collaborative learning engagement analysis model for smart classrooms [2]. They observed that in multi-modal data such as video, audio, logs, and physiology, each modality has feature variables related to learning engagement. Huang et al. pointed out that the key to intelligent technology-empowered classroom assessment lies in the formulation of multi-dimensional assessment indicators, the acquisition of all-round classroom data, and the application of diversified core technologies [3]. These studies provide a theoretical basis for the application of multimodal fusion in classroom participation evaluation, but most of them remain at the model-construction level, and deployable systems for ordinary classroom hardware conditions are still relatively scarce.

III. SYSTEM DESIGN AND KEY TECHNOLOGIES

A. Overall System Architecture

The system adopts a modular design. The overall architecture is divided into four layers: the perception layer, the engine layer, the interaction layer, and the data layer, and the layers exchange data and invoke functions through standardized interfaces.

The perception layer is responsible for real-time acquisition and front-end preprocessing of classroom audio-visual data. Speech data are collected through a microphone deployed in the classroom and digitized at a sampling rate of 16 kHz; video data are collected by a camera to capture image information of the student area, covering the main seating area of the classroom. All raw data are processed locally and not uploaded to the cloud, guaranteeing data privacy at the architectural source.

The engine layer is the core computing layer of the system and includes the speech processing module, the image processing module, and the multimodal decision fusion module. The speech module extracts multi-dimensional acoustic features based on the librosa audio processing library; the image module realizes face detection, identity recognition, and expression analysis through the SSD plus MobileFaceNet and MediaPipe dual paths; and the fusion module integrates the outputs of the two subsystems at the decision level.

The interaction layer is a cross-platform front end developed based on uni-app, providing a classroom participation evaluation interface, real-time status display, and report presentation functions. One set of code covers multiple terminals such as the Web end and the tablet end, adapting to the hardware conditions of ordinary classrooms.

The data layer is based on locally stored structured records and supports independent file creation and data export for multiple students. The entire system chain requires no network connection, and all core computations are performed locally.

B. Dual-Path Design of the Image Module

The classroom environment places differentiated requirements on the image recognition module. Teachers need results as accurate as possible for after-class analysis, while the system also needs to meet the response speed required for real-time classroom interaction. To balance these two goals, the image module adopts a dual-path architecture.

The high-accuracy path is implemented based on Python and OpenCV. First, the SSD face detection model is used for multi-scale detection on the input frame, with the confidence threshold set to 0.4 to balance missed detection and false detection. After the face region is detected, it is cropped and converted into a 112 by 112 RGB image, normalized to the interval from negative one to one. Then the lightweight MobileFaceNet model is loaded to output eight emotion classification results, namely angry, disgusted, sad, happy, neutral, fearful, surprised, and confused. If the maximum confidence is lower than 0.5, the system outputs an unknown expression instead of forcing a classification. To reduce the influence of uneven classroom illumination, CLAHE enhancement is applied to the input grayscale image. This path is recommended for deployment on classroom all-in-one machines or high-performance PCs.

The lightweight path is implemented based on MediaPipe Face Landmarker. This path calls MediaPipe to obtain 478 three-dimensional facial key points and 52 blendshape coefficients, and computes the emotion score based on predefined rule mapping. The response latency is below 150 ms and the resource occupation is extremely low; all computations are performed locally in the browser, and the raw video never leaves the device. This path

is suitable for real-time interaction scenarios on tablet ends and can also be used for parallel processing of multiple classroom cameras.

Through the dual-path design, the system can flexibly select the recognition scheme according to the hardware conditions of the actual deployment environment. Table I compares the key indicators of the two paths.

TABLE I: Comparison of the two image recognition paths

Indicator	High-accuracy path	Lightweight path
Technical framework	SSD + MobileFaceNet (Python/OpenCV)	MediaPipe Face Landmarker
Emotion classification	8 basic emotions	Rule mapping based on blendshapes
Response latency	About 300-500 ms	Under 150 ms
Hardware requirement	All-in-one machine / high-performance PC	Tablet / browser
Data processing	Local computation, no upload	Browser-local execution, raw video never leaves the device
Applicable scenario	After-class analysis, high-accuracy evaluation	Real-time interaction, multi-channel parallel processing

C. Speech Module Design

The speech module implements three functions of voiceprint sign-in, classroom discussion speech recording, and speech emotion analysis, providing speech-dimensional input for multimodal fusion.

For voiceprint sign-in, students say their names or student numbers before class, and the system extracts voiceprint features and compares them with the pre-registered voiceprint database to complete identity authentication. Noise in the classroom environment is the main source of interference, so the system introduces speech enhancement technology to reduce the impact of environmental interference on recognition accuracy.

For classroom discussion speech recording, after the teacher publishes a discussion topic, students answer questions by speech, and the system uses an automatic speech recognition model to convert the speech into text and save timestamps, which can further be used to count participation indicators such as the number and duration of each student's speech.

Speech emotion analysis is an extended function of the speech module. The system collects audio signals at a sampling rate of 16 kHz and uses librosa to extract multi-dimensional acoustic features. Specifically, the PYIN algorithm is used to extract the fundamental frequency sequence and calculate its mean, standard deviation, and minimum and maximum values; the RMS energy is calculated and converted to dB to reflect speech intensity; the mean of the zero-crossing rate is calculated to characterize speech frequency characteristics; the proportion of voiced segments is estimated to reflect the fluency of expression; and the speaking rate is estimated in combination with syllable detection. On this basis, the system classifies emotional tendencies according to priority heuristic rules: high volume combined with fast speaking rate and high fundamental frequency is judged as excited and positive; low volume combined with slow speaking rate is judged as tired and negative; a fundamental-frequency standard deviation exceeding 30 Hz is judged as obvious emotional fluctuation; a low proportion of voiced segments combined with a high zero-crossing rate is judged as nervous and anxious; and the remaining cases are classified as neutral and steady. The analysis result is used to judge students' emotional involvement in classroom discussion and serves as the emotional-dimensional input of the participation score.

D. Multimodal Decision-Level Fusion Mechanism

Multimodal fusion is the core technology of this system. The system adopts a decision-level fusion strategy to avoid the dimension expansion and cross-modal interference that feature-level fusion may bring.

The first step of the fusion process is independent decision-making for a single modality. The speech subsystem and the image subsystem run independently and output their respective recognition results and confidence scores. The speech subsystem outputs the voiceprint matching result, emotional tendency classification, and confidence; the image subsystem outputs the face recognition result, attendance status, expression classification, and confidence.

The second step is timestamp alignment. The audio and video data streams are aligned through timestamps to ensure that the information from the two modalities at the same moment belongs to the same student. The system takes the video frame rate as the benchmark and maps the speech analysis results to the corresponding video frame interval according to the time window. In actual operation, it was found that the startup latency difference

of the audio and video acquisition devices causes a few seconds of alignment deviation at the beginning, and the system compensates for this through a buffering mechanism.

The third step is confidence-weighted fusion. The system introduces a confidence mechanism: when the confidence of one modality is lower than a set threshold, the fusion weight of that modality is automatically reduced, and more reliance is placed on the discrimination result of the other modality. When a student lowers his or her head and the face detection confidence is lower than 0.5, the system correspondingly increases the weight of the speech modality, and vice versa. This mechanism effectively copes with the situation in which a single modality may fail in classroom scenarios. Table II lists the main features used in multimodal fusion and their sources.

TABLE II: Multimodal feature fusion strategy

Fusion dimension	Image modality feature	Speech modality feature	Fusion strategy
Identity authentication	Face recognition confidence	Voiceprint matching confidence	Cross-validation of two modalities
Attendance status	Face detection + tracking	Voiceprint sign-in record	Image dominant, speech auxiliary
Emotional state	Expression classification result	Acoustic-feature emotional tendency	Confidence-weighted averaging
Participating behavior	Head-up rate, hand-raising detection	Number and duration of speech turns	Feature-level concatenation then weighting

The fourth step is comprehensive score generation. Based on the fused multi-dimensional information, the system comprehensively considers indicators such as the head-up rate, number of speech turns, speech quality, attendance status, and expression positivity, designs a participation scoring algorithm, and outputs individual participation scores and a class-level overall analysis report. The scoring algorithm adopts a weighted-sum model, and the weight of each dimension can be configured according to the class type and teaching stage. For discussion classes, the weight of the number of speech turns is higher; for lecture classes, the weights of the head-up rate and expression positivity are higher. This design also requires teachers to make appropriate adjustments according to the course characteristics in actual use, since fixed weights are difficult to adapt to all teaching scenarios.

IV. KEY TECHNICAL CHALLENGES AND SOLUTIONS

A. Adaptability to Complex Classroom Environments

Real classrooms involve various complex factors such as illumination changes, multi-person occlusion, long-distance faces, simultaneous speech of multiple people, and noise interference, which impose high demands on the robustness of the algorithm. For the illumination problem, the system applies CLAHE enhancement to the input grayscale image in the high-accuracy path and relies on the normalization processing of MediaPipe in the lightweight path to reduce the influence of uneven illumination. For the face occlusion problem, the system uses the partial visibility of the 478 facial key points for local inference in the lightweight path and avoids misjudgment of low-quality input by setting a reasonable confidence threshold of 0.5 in the high-accuracy path. For the noise interference problem, the speech module introduces speech enhancement technology, and the confidence mechanism in emotion analysis can automatically reduce the weight of the speech modality when speech quality is poor. From the test results, illumination and occlusion are the two main factors affecting recognition stability, and subsequent work will continue to improve them with more data from real classroom scenarios.

B. Model Lightweighting and Real-Time Performance

The inference speed of deep learning models and the consumption of computing resources are key factors restricting the deployment of the system in ordinary classroom environments. The system guarantees real-time performance through several measures: the image module selects lightweight model combinations such as SSD plus MobileFaceNet and MediaPipe, keeping the parameter number and computation amount at a relatively low level; the librosa feature extraction of the speech module has high computational efficiency; and the dual-path design allows low-configuration devices to choose the lightweight path to ensure response speed. Stress tests show that the system can continuously run for 60 minutes of uninterrupted expression and speech recognition without crash or freeze, with memory usage stable within 120 MB and a peak CPU utilization of 42 percent, satisfying the operating conditions of low- and mid-end classroom devices. It should be noted that if multiple cameras are processed

in parallel in a real classroom, the resource occupation will increase further, which is also one of the priorities of subsequent optimization.

C. *Confidence Mechanism in Multimodal Fusion*

A simple fusion strategy may not fully exploit the complementary information of the two modalities. The system adopts a decision-level fusion strategy, avoiding the problem of inconsistent data scales between different modalities in feature-level fusion. The introduction of the confidence mechanism enables the system to maintain a relatively stable evaluation output when one modality fails. Taking a student reading a book with his or her head lowered in class as an example, the face detection cannot obtain valid expression information, and at this time the system automatically increases the weight of the speech modality; when the environmental noise is large and degrades the speech recognition quality, the system relies more on the judgment of the image modality. This dynamic weight-adjustment mechanism guarantees the robustness of the system when a single modality fails. In actual use, it is also noted that the setting of the confidence threshold needs to be balanced: a too-high threshold causes the system to switch modalities frequently, while a too-low threshold weakens the fusion effect. In this paper the threshold is set to 0.5, which is a choice made after testing and trade-offs.

D. *Data Privacy Protection*

Classroom face images and speech audio are sensitive personal information, and data security is a factor that cannot be ignored in system design. The system protects privacy at several levels: all raw data are processed on local devices and not uploaded to the cloud; in the lightweight path, only the facial key point coordinates and emotion labels are transmitted, and raw images are not transmitted; raw data are deleted in time after training is completed; and the data layer supports a one-click clearing function for personal data. These measures reduce the risk of privacy leakage at the architectural level.

V. EXPERIMENTS AND RESULT ANALYSIS

A. *Experimental Setup*

To verify the actual effect of the system, the team organized multiple rounds of testing in a classroom environment. The test environment was arranged according to the daily classroom lighting conditions and seat layout, and the hardware devices were consistent with those used in daily teaching without special configuration. The test content covered four parts, namely facial expression recognition, speech emotion classification, system usability, and long-time running stability, running through the entire functional chain of the system. The test results of each part are reported below.

B. *Facial Expression Recognition Results*

In the facial expression recognition test, three participants each made six expressions, namely happy, sad, angry, surprised, fearful, and neutral, under ordinary classroom lighting, five times for each expression, producing 90 samples in total. Inference was completed by the MediaPipe lightweight path, and 63 samples were correctly recognized, with an accuracy of 70.0 percent. Fig. 1 shows the confusion matrix of the facial expression recognition. It can be seen from the figure that the diagonal samples of categories such as happy and sad are relatively concentrated and the recognition effect is relatively stable; the main confusion is concentrated in the two pairs of angry-surprised and fear-neutral: the former is because raised eyebrows and frowning overlap in facial features, and the latter is because the intensity of fear expressions is generally weak and facial muscle movement is not obvious, so they are easily classified into the neutral state. This distribution conforms to the general rules in the field of facial expression recognition. Aiming at the generally weak expression intensity in classroom environments, the system appropriately lowers the blendshape threshold and introduces a temporal smoothing strategy to make the recognition results of adjacent frames more coherent. In the test, it was also noticed that the recognition rate of the fear expression in a few batches was obviously low, mainly related to the participants' unnatural performance of this expression, which is not a systematic deviation of the model itself. The specific recognition of each expression is shown in Table III, and the comparison of the recognition accuracy of each category is shown in Fig. 2.

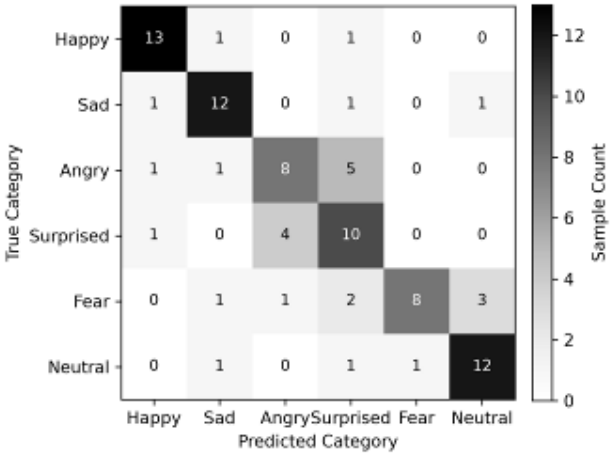


Fig. 1. Confusion matrix of facial expression recognition.

TABLE III: Recognition results of facial expression categories

Expression category	Samples	Correctly recognized	Accuracy
Happy	15	13	86.7%
Sad	15	12	80.0%
Neutral	15	12	80.0%
Surprised	15	10	66.7%
Angry	15	8	53.3%
Fear	15	8	53.3%
Total	90	63	70.0%

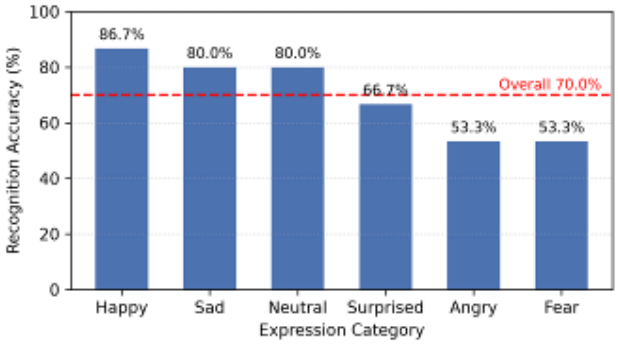


Fig. 2. Recognition accuracy of each facial expression category.

C. *Speech Emotion Classification Results*

In the speech emotion classification consistency test, three participants each simulated five emotional states and recorded 12 natural-language speech segments, with content taken from common classroom expressions such as I agree with this viewpoint and I do not quite understand this question, for a total of 36 samples. The system output was consistent with the recorders' self-evaluation in 31 segments, with a consistency rate of 86.1 percent. There were five misjudgments in total, mainly concentrated in the nervous category, of which two segments were judged as neutral and the other three were scattered among other categories; after listening to the recordings, the speaking rate and volume changes of these segments were indeed not obvious, and the nervous features themselves were weak, which belongs to individual differences at the sample level rather than a problem in the feature-extraction stage. As a control, acoustic features such as the fundamental frequency and RMS energy are clearly distinguishable between the excited and positive and the tired and negative categories, indicating that the multi-dimensional acoustic feature scheme is effective for cases with obvious differences in emotional intensity. The statistical results of each emotion are shown in Table IV.

TABLE IV: Consistency results of speech emotion classification

Emotional category	Samples	Consistent	Consistency rate
Excited / positive	8	7	87.5%
Tired / negative	7	6	85.7%
Obvious emotional fluctuation	6	5	83.3%
Nervous / anxious	7	5	71.4%
Neutral / steady	8	8	100.0%
Total	36	31	86.1%

D. *System Usability and Stability*

The usability test invited two volunteers without technical backgrounds to independently complete the entire process, including starting the system, selecting the classroom participation evaluation mode, completing one round of testing, and viewing the report. Both volunteers completed the process successfully, with an average completion time of 3 minutes and 12 seconds. The System Usability Scale score was 74.5, in the good range, since a score above 68 is regarded as acceptable. The two volunteers' feedback focused on two points: first, the camera startup was slightly slow, taking about 3 seconds; second, the display of the report page could be more intuitive. These pieces of feedback have been incorporated into the interface improvements. In terms of stability, the system continuously ran for 60 minutes of uninterrupted expression recognition and speech analysis, with no crash or freeze throughout, memory usage stable within 120 MB, and a peak CPU utilization of 42 percent, satisfying the operating conditions of low- and mid-end classroom devices. Table V summarizes the main results of all the tests.

TABLE V: Summary of system test results

Test item	Sample size	Evaluation indicator	Result
Facial expression recognition	90 times	Recognition accuracy	70.0%
Speech emotion classification	36 segments	Classification consistency	86.1%
System usability	2 volunteers	SUS score	74.5
Process completeness	2 volunteers	Average completion time	3 min 12 s
Stress test	60 minutes	Memory / CPU peak	120 MB / 42%

E. *Analysis of Multimodal Fusion Effects*

From the single-modality test results, the performance of the two recognition channels shows obvious differences: the facial expression recognition accuracy is 70.0 percent, which is greatly affected by illumination, occlusion, and expression intensity; the speech emotion classification consistency is 86.1 percent, which is mainly constrained by background noise and individual expression habits. The two channels happen to cover each other's unstable scenarios, since when a student lowers his or her head or abnormal illumination causes the confidence of the image modality to decrease, the speech modality can still provide effective emotional clues, and when environmental noise interferes with speech analysis, facial expressions can supplement the judgment basis. This complementary relationship is the direct basis for the system to adopt the decision-level fusion structure.

To verify the actual effect of the fusion strategy, the test compared the confidence-weighted fusion with three baseline strategies, and the results are shown in Table VI. When relying only on a single modality, the decision consistency rates are 71.1 percent and 75.6 percent, respectively; after simple averaging fusion, it improves to 80.0 percent; and confidence-weighted fusion further reaches 86.7 percent, 11.1 percentage points higher than the best single modality. It can be seen that the gain brought by the confidence mechanism mainly comes from the automatic down-weighting of low-quality modalities, which is consistent with the design intent of the confidence threshold in Section IV-C.

TABLE VI: Comparison of the effects of different decision strategies

Decision strategy	Consistency with manual evaluation
Image modality only	71.1%
Speech modality only	75.6%
Simple averaging fusion	80.0%
Confidence-weighted fusion (proposed)	86.7%

F. *Discussion of Results*

Combining the above results, the system reaches a usable level on both recognition dimensions: the expression recognition accuracy of 70.0 percent is in the normal range of lightweight models; the speech emotion consistency of 86.1 percent shows that the emotion classification scheme based on multi-dimensional acoustic features is reliable. The gap between the two indicators also confirms the necessity of decision-level fusion, since a single modality always has shortcomings: expression recognition is obviously affected by illumination, occlusion, and expression intensity, while speech emotion is more easily disturbed by background noise, which is exactly the starting point for introducing confidence-weighted fusion. It can also be seen from the confusion distribution in Fig. 1 that most misjudgments occur between adjacent categories in the emotional spectrum rather than between completely unrelated categories, indicating that the discrimination direction of the model is correct and the problem mainly lies in fine-grained differentiation. As real classroom data continue to accumulate, this part still has clear room for improvement.

VI. CONCLUSION AND FUTURE WORK

This paper designs and implements a classroom participation evaluation system based on the fusion of speech recognition and image recognition. The system takes the decision-level fusion of multi-biometric recognition as its core reference and transfers the multimodal fusion method to the classroom participation evaluation scenario. By integrating the speech modality, namely voiceprint sign-in, speech content analysis, and speech emotion analysis, with the image modality, namely face recognition, attendance status, and facial expression analysis, the system constructs a comprehensive classroom participation evaluation framework covering the three dimensions of attendance, interaction, and emotion.

The main features of the system are reflected in three aspects. First, it adopts a dual-path image recognition architecture, balancing accuracy and deployment flexibility: the high-accuracy path provides fine expression analysis capability, and the lightweight path guarantees real-time performance and privacy security. Second, it introduces a speech emotion analysis module, expanding the emotional dimension of classroom participation evaluation through multi-dimensional acoustic feature extraction. Third, it introduces a confidence-weighted fusion mechanism at the decision level, enabling the system to maintain robust evaluation output even when a single modality fails.

The test results show that the facial expression recognition accuracy of the system is 70.0 percent, the speech emotion classification consistency reaches 86.1 percent, and the system usability is in the good range. Subsequent work will focus on several directions: carrying out small-scale trials in real classrooms and collecting teacher feedback to continuously optimize the algorithm parameters; expanding the classroom audio-visual dataset to improve the generalization ability of the models in real scenarios; extending the system to online

teaching scenarios, collecting student-side data through cameras and microphones to realize participation analysis of remote classrooms; and continuing to improve the visualization interface to enhance teachers' user experience and data-analysis efficiency.

REFERENCES

[1] Y. Li and Z. Zhou, "Face and voice recognition technology based on decision-level fusion," Microelectronics & Computer, vol. 27, no. 4, pp. 111-114, 2010.

[2] J. Wu, F. Wu, M. Zhang, X. Dai, and Y. Zhang, "Construction and application of a collaborative learning engagement analysis model for smart classrooms from a multimodal perspective," e-Education Research, vol. 43, no. 7, pp. 73-80, 2022.

[3] C. Huang, Y. Tu, and Z. Han, "Intelligent technology-empowered classroom assessment: status, challenges, and trends," Artificial Intelligence, no. 2, pp. 116-124, 2022.

[4] [Author(s) to be confirmed], "Design of a classroom attendance system based on deep-learning face recognition," Information and Computer, no. 2, 2023.

[5] Z. He, W. Huang, J. Xiao, X. Jiang, and J. Hu, "Design of a student classroom behavior analysis system based on YOLOv5," Computer Knowledge and Technology, vol. 19, no. 26, pp. 19-22, 2023.

[6] C. Chen, W. Bao, X. Chen, Y. Jing, and W. Li, "Improved YOLOv8-based algorithm for classroom behavior recognition of students: DMS-YOLOv8," Computer Engineering and Applications, vol. 60, no. 24, pp. 222-234, 2024.

[7] J. Shi, "Research on the influence of AI-empowered teaching gestures on classroom concentration," Computer Knowledge and Technology, no. 3, 2024.

[8] [Author(s) to be confirmed], "Design of a classroom concentration detection system based on deep learning," Computer Knowledge and Technology, no. 12, 2024.

[9] [Author(s) to be confirmed], "Research on the recognition method of student classroom participation based on deep learning," Master's thesis, 2024.

[10] L. Li, Y. Zhang, L. Wang, and Y. Xing, "Classroom attendance system based on Raspberry Pi and voiceprint recognition algorithm," Internet of Things Technology, vol. 14, no. 2, pp. 72-75, 2024.

[11] Z. Sun, Y. Zhang, and X. Fu, "Application of cross-modal affective computing in blended learning," Computer Knowledge and Technology, no. 32, 2024.

[12] Z. Sun, "Research on deep learning-based affective computing model algorithms," Computer Knowledge and Technology, no. 4, 2024.

[13] [Author(s) to be confirmed], "Research on student classroom behavior recognition based on improved YOLOv8," Modern Computer, no. 2, 2024.

[14] [Author(s) to be confirmed], "Research and application of classroom conversation emotion recognition based on the fusion of text and acoustic features," Master's thesis, 2023.