Computers, Materials & Continua

Tech Science Press

https://doi.org/10.32604/cmc.2026.081460

ARTICLE

Improvement of Emotion Detection by Fusing Speech and Image Based on CNN with Temporal Models

Shing-Tai Pan*, Yi-Zhen Huang and Zhi-Qing Chen

Department of Computer Science and Information Engineering, National University of Kaohsiung, Kaohsiung, Taiwan

*Corresponding Author: Shing-Tai Pan. Email: stpan@nuk.edu.tw

Received: 08 March 2026; Accepted: 21 May 2026; Published: 23 July 2026

ABSTRACT: This paper proposes a multimodal fusion framework that integrates speech and visual features to enhance the accuracy of emotion recognition. The principal contribution lies in extending the visual component from single-image to multi-image emotion recognition. Specifically, the proposed framework employs an InceptionV3 Convolutional Neural Network (CNN)-based architecture to extract features from multiple facial images representing the speaker's expressions throughout an utterance. These features are concatenated into a single vector and subsequently processed by Long Short-Term Memory (LSTM) or Hidden Markov Model (HMM) for temporal modeling. For the speech modality, Mel-Frequency Cepstral Coefficients (MFCC) or filter bank features are extracted from processed audio signals and fed into a hybrid CNN–time-series model. The two modalities are then integrated through model-level and decision-level fusion strategies. Since recognition accuracy tends to degrade as the number of utterances and speakers increases, the Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS), which contains a moderate number of sentences and speakers, is adopted in this study. Experimental results demonstrate that the proposed multi-image approach improves recognition accuracy from 91% to 96% compared with the single-image baseline, and that the multimodal fusion framework consistently outperforms its single-modal counterpart.

KEYWORDS: Speech emotion recognition; consecutive facial image emotion recognition; convolutional neural network (CNN); long short-term memory (LSTM); hidden markov model (HMM); support vector machine (SVM)

1 Introduction

1.1 Research Motivation and Objectives

Emotions constitute one of the most informative non-verbal channels of communication in both humans and animals, are typically manifested through physiological responses and bodily expressions. They are commonly classified into seven categories: happiness, disgust, fear, surprise, neutrality, anger, and sadness. In 1978, Ekman and Friesen [1] proposed the Facial Action Coding System (FACS), a widely adopted framework for facial expression analysis. FACS is based on Facial Action Units (FAUs), which represent fundamental facial muscle movements such as eyebrow raises and mouth corner movements.

With the rapid advancement of artificial intelligence, human-like machines have been deployed to substitute for human labor in service context, including unmanned convenience stores, caregiving robots, customer service agents, and hotel concierge robots. Prolonged interaction with such machines, however, often induces discomfort in users, as communication tends to be rigid and confined to standardized operations and message-delivery patterns. In contrast, human interaction relies heavily on emotional cues, which enable interlocutors to respond appropriately and establish meaningful social bonds. Endowing

machines with the ability to perceive and respond to human emotions could therefore substantially narrow the affective gap between humans and machines. According to psychologist Mehrabian [2], only 7% of interpersonal communication is conveyed through verbal content, whereas 38% is attributed to vocal cues and 55% to facial expressions. Motivated by this finding, this thesis investigates how speech-based and facial-expression-based emotion recognition can be jointly leveraged to enable more natural and emotionally aware human–machine interaction.

Currently, the most widely adopted approaches to emotion recognition fall into three categories: facial expression recognition, speech emotion recognition, and physiological signal-based recognition. This thesis focuses on the first two modalities and proposes to integrate temporally continuous facial expression features with speech emotion recognition results through a Support Vector Machine (SVM) [2]. By combining the outputs of these two recognition modules, the proposed framework is expected to achieve higher overall accuracy in emotion recognition and thereby contribute to more natural human–machine interaction [3].

1.2 Research Structure

The emotion recognition framework proposed in this study consists of three main stages.

(1) Speech Emotion Recognition: The first step extracts acoustic features from the speech signal. This study adopts MFCC [4] and filter bank features [5,6]. Both feature extraction methods follow similar procedures: preprocessing the raw signal and converting it from the time domain to the frequency domain using the Fast Fourier Transform (FFT). The key difference lies in the final step, where MFCC applies **a** Discrete Cosine Transform (DCT) to compress the feature representation. After feature extraction, the extracted features are fed into hybrid model—CLDNN which combines CNN [7,8], LSTM [8,9], and a Deep Neural Network (DNN)—or a Hidden Markov Model (HMM) [10] for training and testing [11]. The overall process is illustrated in Fig. 1.

Figure 1: Block diagram of the speech emotion recognition system.

(2) Sequential Facial Image Emotion Recognition: The process begins by extracting key frames from a video in two steps. First, a representative frame is selected to characterize the emotion conveyed in the video. Second, the video is uniformly sampled to obtain N frames. Each extracted frame is then preprocessed. The recognition phase is likewise divided into two steps. In the first step, the representative image is used to train the InceptionV3 model from which the model weights are obtained. In the second step, the N sampled frames are tested using the trained InceptionV3 model, and the resulting outputs are fed into time-series models—namely LSTM and HMM [4,12]—for sequential emotion training and testing.

The overall procedure is illustrated in Fig. 2.

(3) Fusion of Multimodal Emotion Recognition Results: In this stage, the outputs from the speech emotion recognition (Step 1) and the sequential image-based recognition (Step 2) are combined. Two fusion strategies are applied: decision-level fusion and model-level fusion.

Specifically, decision-level fusion employs an SVM to integrate the classification results from both modalities, whereas model-level fusion utilizes a Concatenate layer to merge the feature vectors prior to the final classification. The final emotion recognition result is then obtained from the fused output. The detailed fusion processes are illustrated in Figs. 3 and 4.

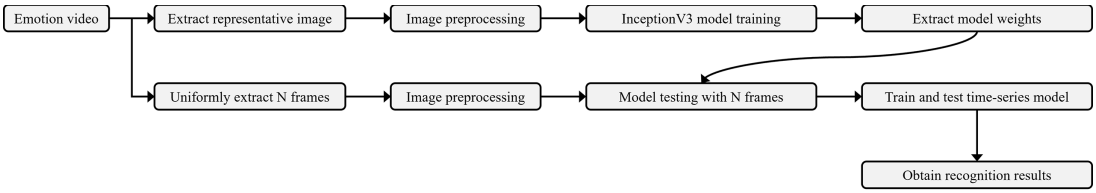

Figure 2: Sequential image-based emotion recognition process flow.

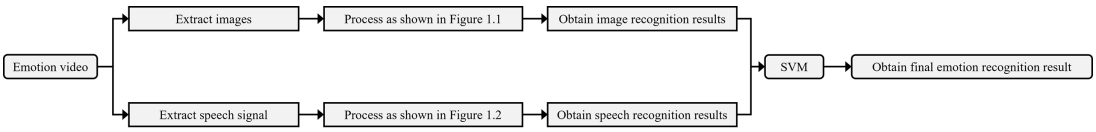

Figure 3: Emotion recognition process via decision-level fusion of speech and sequential images.

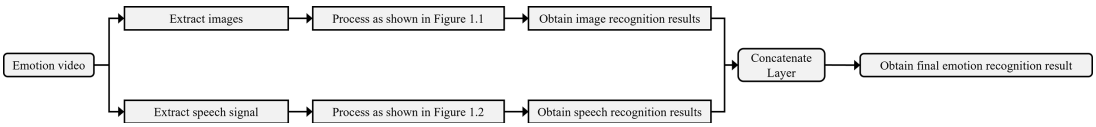

Figure 4: Emotion recognition process via model-level fusion of speech and sequential images.

The experiments in this study are conducted on the RAVDESS dataset [13]

The main contributions of this work are summarized as follows:

(i) Transition from single-image to sequential-image emotion recognition. Most existing audio-visual emotion recognition systems extract a single representative frame per utterance and discard the temporal dynamics of facial expressions during speech. This work systematically replaces the single-frame visual branch with a multi-frame sequential branch under a unified multimodal fusion framework, and empirically demonstrates the resulting accuracy improvement from 91% to 96% on the RAVDESS dataset.

(ii) Empirical analysis of the optimal frame-sampling number. A parameter study is conducted to evaluate the recognition accuracy across different values of N (the number of sampled frames per video). The results indicate that performance saturates around N = 40, identifying this value as the optimal trade-off between recognition accuracy and computational efficiency for the RAVDESS dataset.

(iii) Innovative improvement of computational efficiency by combining the HMM model with CNNs for temporal data recognition, and a comparative analysis of HMM and LSTM as temporal backbones. Within a unified multimodal framework, this work provides a side-by-side comparison of HMM and LSTM along two complementary dimensions—recognition accuracy and computational efficiency—accompanied by a mechanism-level analysis explaining the observed differences.

(iv) Comparative analysis of decision-level and model-level fusion strategies. A controlled comparison of these two fusion paradigms is provided within an otherwise identical pipeline, with mechanism-level interpretation of why model-level concatenation outperforms SVM-based decision fusion in both mean accuracy and variance.

2 Related Work

Previous studies on emotion recognition typically follow three approaches: (1) facial emotion recognition using extracted facial features, (2) speech emotion recognition using signal features like MFCC or

filter banks, and (3) multimodal fusion, combining facial and speech features. Some related works are discussed below.

2.1 Speech Emotion Recognition

Fahad et al. [14] used a DNN-HMM model for speech emotion recognition with preprocessed features, including MFCC, epoch-based features [15], and their combination. Their results showed that combining MFCC and epoch-based features yielded the best performance. Wu et al. [16] used capsule networks (CapsNet) [17] to capture spatial relationships in speech features. They introduced a pooling strategy to build hierarchical representations of pronunciation features, which were processed by a recurrent neural network (RNN) to improve speech emotion recognition accuracy. Liu et al. [18] proposed a model combining Time-Frequency CNN [19] and CapsNet for speech spectrograms. The TFCNN captures local features while CapsNet captures global features, and the resulting representations are fused through fully connected layers for emotion classification.

While these works primarily focus on hand-crafted spectral features (e.g., MFCC) combined with CNN- or RNN-based models, recent studies have increasingly shifted toward self-supervised learning (SSL) and transformer-based architectures. Self-supervised pre-trained models such as Wav2Vec 2.0 and HuBERT have been shown to outperform conventional spectral-feature pipelines in speech emotion recognition tasks; Jafarzadeh et al. [20] reported substantial gains on the RAVDESS dataset—the same corpus adopted in this paper—by leveraging fine-tuned HuBERT representations. Dabbabi and Mars [21] further demonstrated that the lightweight Distil HuBERT achieved 87.01% accuracy on RAVDESS while substantially reducing model size, indicating the practical value of SSL-based representations for resource-constrained scenarios.

2.2 Facial Emotion Recognition

Ma et al. [22] improved emotion recognition using SVM and Deep Boltzmann Machines (DBM) with feature-level, decision-level [23], and model-level fusion strategies. Zhang et al. [24] extracted facial features via the Haar cascade classifier [25] and Ada Boost, then trained a CNN. Their method outperformed R-CNN [26] and Faster R-CNN [27] in recognition accuracy, though at the cost of inference speed. Purakkadavath and Chacko [28] used an InceptionV3-LSTM model to detect fatigue from 68 facial landmarks identified by Gradient Boosting Decision Trees (GBDT) with InceptionV3 extracting sequential features for LSTM-based temporal classification.

While these studies demonstrate the effectiveness of CNN-based architectures for facial emotion recognition, recent research has increasingly adopted Vision Transformer (ViT)–based models, which leverage self-attention mechanisms to capture global spatial dependencies that conventional CNNs often overlook. Tian et al. [29] proposed HLA-ViT, a dual-stream architecture that integrates hybrid local attention with global contextual modeling, demonstrating improved robustness against occlusion and pose variation. Kus et al. [30] further provided a systematic review of ViT-based and explainable AI approaches in multimodal facial expression recognition, identifying ViT as a dominant trend that has progressively replaced CNN backbones in recent literature.

2.3 Multimodal Emotion Recognition Combining Speech and Facial Features

Liu [31] proposed a multimodal fusion using a Coupled HMM (CHMM) [32], improved to a Semi-Coupled HMM (Semi-CHMM) [33], aligning facial and speech state sequences and fusing them via error-weighted classification for emotion recognition. Chen [8] employed decision-level fusion via a lookup table to combine speech and facial recognition outputs. Speech features were extracted using MFCC and processed by a CNN with an LSTM replacing the fully connected layer; facial features were extractedfrom

manually selected frames and processed by a CNN. Ghaleb et al. [34] proposed MERML, which aggregates video and speech features via Fisher Vectors [35] into sequence representations. These representations are then used to train an SVM with a Radial Basis Function (RBF) kernel for the final classification.

Overall, these studies show that most multimodal fusion strategies—both decision-level and model-level—consistently outperform unimodal methods, as evidenced by both quantitative experiments and visualization analyses. More recent multimodal frameworks have moved beyond conventional fusion paradigms toward transformer-driven cross-modal architectures. Yi et al. [36] proposed HyFusER, a hybrid multimodal transformer that employs dual cross-modal attention to dynamically align acoustic and visual representations, achieving state-of-the-art performance on standard MER benchmarks. Comprehensive surveys by Ramaswamy and Palaniswamy [37] and Lian et al. [38] further consolidate the observation that transformer-based fusion, attention-driven gating, and self-supervised pre-trained encoders are progressively redefining the design space of MER, suggesting promising directions in which the framework proposed in this thesis can be extended.

3 Methods

This study uses a SVM to integrate classification results from multiple deep learning models for emotion recognition, based on speech and sequential facial expressions. The system has five main components, shown in Fig. 5:

1. Feature Extraction
2. CNNs
3. Temporal Sequence Models
4. Facial and Speech Model Architectures
5. Bimodal Emotion Recognition Model for Speech and Visual Modalities

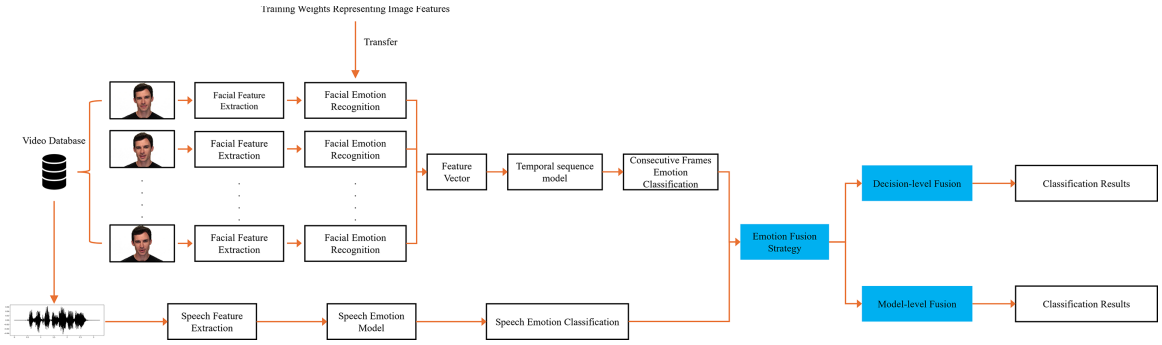

Figure 5: The proposed architecture for multimodal emotion recognition based on speech and continuous visual input is illustrated in this figure.

3.1 Feature Extraction

This section introduces the feature extraction methods proposed in this study, focusing on the extraction of speech and facial emotion features.

3.1.1 Speech Feature Extraction

Speech data in this study come from emotional video clips. Emotions are inferred from prosodic, acoustic, and spectral cues, but recordings are often affected by environmental or device noise, making noise reduction essential before feature extraction. Here, filter banks and MFCCs are used as primary features.

MFCCs compress spectral information via filter banks, modeling human auditory perception. The extraction steps are shown in Fig. 6.

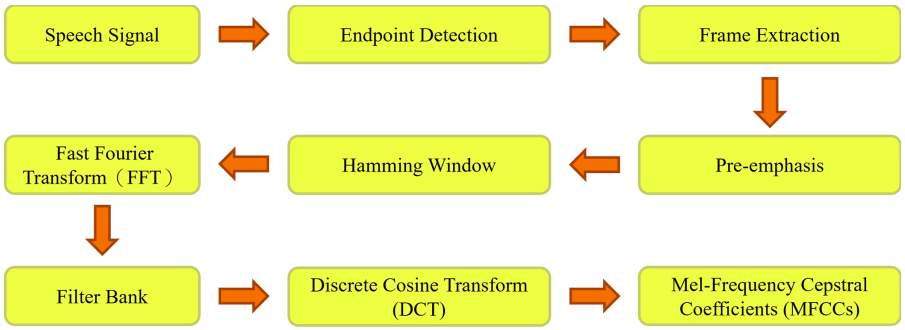

Figure 6: The preprocessing pipeline of the speech signal.

3.1.2 Facial Feature Extraction

In this study, emotional video clips are processed by selecting a representative image and a sequence of continuous facial frames. The representative image trains a facial expression model, whose weights are then used to test each frame in the sequence, producing video feature vectors. A time-series model is then applied to capture temporal relationships between frames, enabling overall emotion recognition across the sequence. For facial region extraction, the Haar Cascade Classifier is used to locate the Region of Interest (ROI), mainly the facial area. The dlib Python library then extracts 68 facial landmarks, each with (x, y) coordinates for specific facial structures, as shown in Fig. 7.

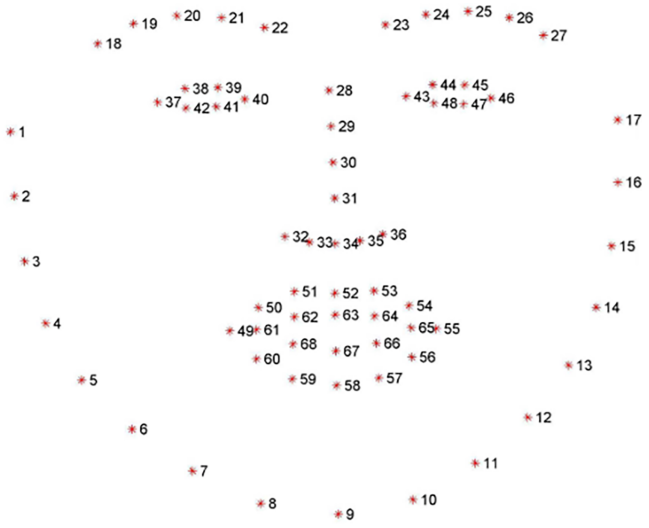

Figure 7: 68 facial landmark points.

These landmarks are predicted using a shape predictor trained on the iBUG 300-W dataset [32]. Non-facial regions are then cropped, and the facial images are resized to 314 × 314 pixels and normalized. The complete facial feature extraction process is shown in Fig. 8.

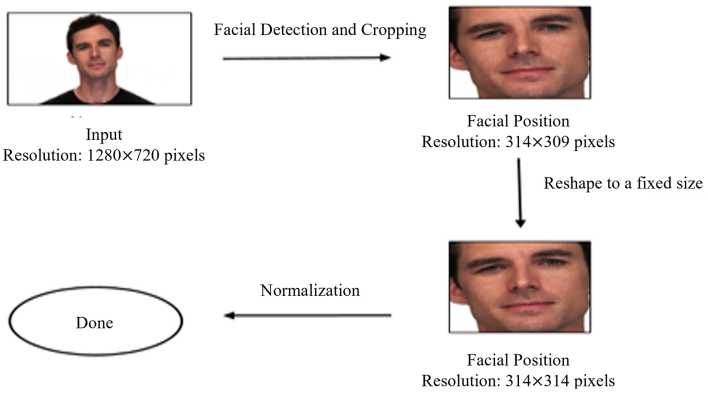

Figure 8: Flowchart of the facial extraction process.

3.2 Feature Emotion Image Recognition Model

This study uses temporal models, including CNN and InceptionV3 [39], to train on representative emotional images from videos. These models learn emotion-specific features, and the trained weights are then applied to continuous image sequences as a preprocessing step for sequential emotion analysis.

3.2.1 Convolutional Neural Network (CNN)

A CNN typically includes convolutional layers, average or max pooling layers [40], and fully connected layers. CNNs extract features via local receptive fields, capturing spatial patterns, and stacking layers enables learning of increasingly abstract, hierarchical representations. In this study, CNNs process 314 × 314 pixel facial regions to extract features from images and speech-derived data. Convolutional layers capture local features, and max pooling retains the most prominent values, reducing computation and highlighting relevant activations.

3.2.2 InceptionV3

This study uses a relatively shallow InceptionV3 model [39] for single-image emotion recognition. Its architecture, shown in Figs. 9 and 10 [28], replaces large convolutional filters with multiple smaller ones, reducing computation and complexity while increasing non-linearity to support deeper networks and reduce overfitting. InceptionV3 is chosen because deeper models like InceptionV4 or ResNet-152 offer little benefit in sequence-based emotion recognition, where features from multiple images are passed to a time-series model. Its low single-image error rates make it suitable for this study.

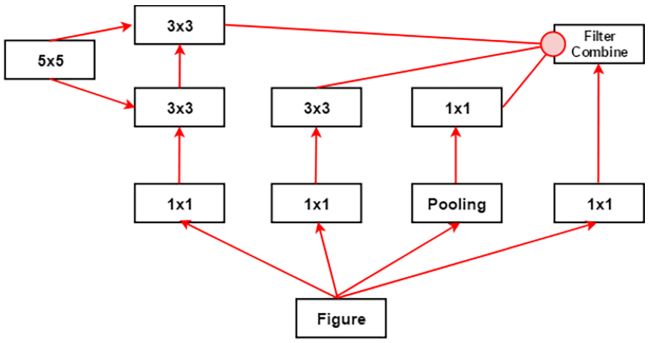

Figure 9: Inception model [29].

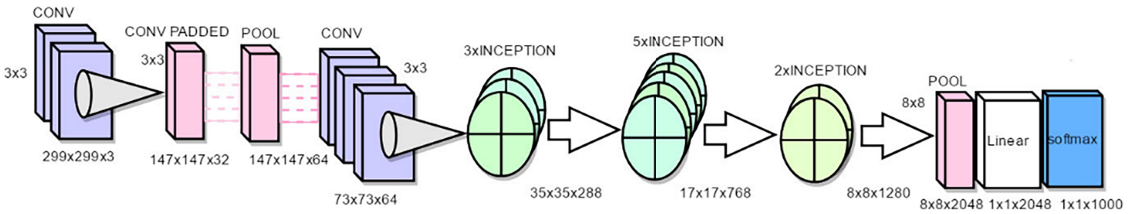

Figure 10: Inception V3 model [29].

3.3 Time-Series Model

This study employs temporal models including HMM [41,42], RNN [43], and LSTM networks. These models process current and past information to predict future states. The following sections detail each model.

3.3.1 Hidden Markov Model (HMM)

In this study, a 40-dimensional feature vector is input to a temporal recognition system using Discrete HMMs (DHMMs) [41,42,44]. HMMs, as probabilistic models, handle variable-length sequences like continuous images or speech with less training time than deep learning methods.

A separate HMM is trained for each emotion. During recognition, the likelihood of the observation sequence is computed for each model, and the highest likelihood determines the classification. Continuous image sequences are treated as observation sequences, with feature changes modeled through hidden state transitions. For each emotion-specific HMM, emission and state transition probabilities are trained, while all models start from the same fixed initial state.

Let the HMM be defined as: $\lambda = \{A, B, \pi, S, V\}$

- $S = \{s_1, s_2, \ldots, s_N\}$: Set of N hidden states
- $V = \{v_1, v_2, \ldots, v_N\}$: Set of M observable symbols (observations)
- $A = \{a_{ij}\}, a_{ij} = P\left(q_t = s_j | q_{t-1} = s_i\right)$: State transition probability matrix
- $B = \{b_j(k)\}, b_j(k) = P\left(o_t = v_k | q_t = s_j\right)$: Observation (emission) probability matrix
- $\pi = \{\pi_i\}, \pi_i = P\left(q_1 = s_1\right), 1 \leq i \leq N$: Initial state probability vector
- $O = \{o_1, o_2, \ldots, o_T\}$: Observation sequence
- $Q = \{q_1, q_2, \ldots, q_T\}$: Hidden state sequence

3.3.2 Long Short-Term Memory (LSTM)

Traditional RNNs struggle with large information sequences and suffer from vanishing or exploding gradients. LSTM networks address these issues with a more complex architecture, including an input gate, forget gate, output gate, and cell state update unit [39], each with separate weights. These gates control whether information is passed through or blocked.

In the context of this study, the LSTM operates directly on continuous-valued probability vectors produced by the InceptionV3 frame-level classifier. For each video, the 40 sequentially sampled frames yield a sequence of 40 emotion probability vectors, which are fed into the LSTM in temporal order. Through its gating mechanism, the LSTM jointly models the gradual transitions between frames—capturing how the speaker's facial expression evolves throughout the utterance—rather than treating each frame as an independent observation. Unlike the Discrete HMM described in Section 3.3.1, which requires a vector-quantization

step to map continuous probability vectors into a finite codebook, the LSTM preserves the original continuous representation, retaining fine-grained intensity information across the 40-frame sequence.

4 Experiments and Results

This section first describes the experimental environment and dataset. We then evaluate the proposed emotion recognition system using three approaches: speech-based, image-based, and multimodal fusion. Finally, results are compared with related works to demonstrate the method's effectiveness.

4.1 Environment

Experiments were conducted on an NVIDIA GeForce RTX 3090 GPU with 32 GB RAM. The environment used Python 3.8.5 within Anaconda3, and experiments were implemented in Jupyter Notebook using TensorFlow-nightly-GPU 2.4.0. The detailed software and hardware specifications used in the experiments are summarized in Table 1.

Table 1: Software and hardware specifications.

Category	Specification
CPU	Intel(R) Core(TM) i7-10700 @ 2.90 GHz
Memory	16 GB × 2 GB
GPU	NVIDIA GeForce RTX 3090
CUDA Cores	10,496
Development Tools	Anaconda3, Jupyter Notebook
Programming Language	Python 3.8.5
Deep Learning Library	TensorFlow-nightly-GPU 2.4.0
Cross-validation	10-fold cross-validation

4.2 Dataset

This study uses the RAVDESS dataset [31] for experiments, containing recordings from 24 professional actors (12 male, 12 female) expressing seven emotions—neutral, happy, sad, angry, fearful, disgust, and surprised—through spoken and sung versions of two sentences. Video clip durations vary, and Table 2 summarizes the dataset characteristics.

Table 2: Summary of the RAVDESS dataset.

Category	Description
Dataset Name	RAVDESS
Language	North American English
Number of Speakers	Male: 12 Female: 12
Number of Emotions	7 types
Expression Mode	Singing, Speaking
Number of Sentences	2
Video Duration	2–5 s
Total Samples	2428

4.3 Speech Recognition Results

The main goal of this experiment is to compare features and models for speech-based emotion recognition. The following sections detail the feature parameters and models used. After preprocessing, two speech features—MFCC and filter banks—were extracted and input into CNN-based deep learning models for emotion recognition. Models were trained for 300 iterations, and their performance was evaluated using K-fold cross-validation. Table 3 shows the results for different feature-model combinations.

Table 3: Recognition accuracy (%) of different feature–model combinations for speech emotion recognition on the RAVDESS dataset.

Method	CNN	CNN-HMM	CLDNN	CNN	CNN-HMM	CLDNN
Feature	**MFCC**	**MFCC**	**MFCC**	**Filter bank**	**Filter bank**	**Filter bank**
Fold	**Accuracy**	**Accuracy**	**Accuracy**	**Accuracy**	**Accuracy**	**Accuracy**
Fold 1	0.568	0.624	0.736	0.742	0.737	0.822
Fold 2	0.584	0.629	0.767	0.716	0.711	0.841
Fold 3	0.539	0.644	0.681	0.757	0.835	0.785
Fold 4	0.609	0.680	0.773	0.762	0.794	0.822
Fold 5	0.625	0.592	0.760	0.742	0.742	0.798
Fold 6	0.584	0.639	0.767	0.726	0.773	0.847
Fold 7	0.662	0.603	0.779	0.747	0.789	0.847
Fold 8	0.625	0.582	0.779	0.762	0.701	0.815
Fold 9	0.615	0.593	0.765	0.731	0.737	0.796
Fold 10	0.603	0.611	0.775	0.694	0.736	0.784
Mean	0.602	0.619	0.758	0.736	0.755	0.816
SD	0.034	0.02	0.03	0.022	0.041	0.034
Variance	0.0012	0.0008	0.0009	0.0005	0.0017	0.0012

This study uses CNN, CNN-HMM, and CLDNN models. Table 4 compares MFCC and filter bank features, showing that filter banks (without DCT compression) achieve higher accuracy. CNN alone performs worse than temporal models like CNN-HMM and CLDNN, with the lowest accuracy of 0.602 using CNN with 12-dimensional MFCC features. In contrast, the highest accuracy of 0.816 was achieved using CLDNN with 40-dimensional filter bank features. The observation that filter bank features outperform MFCCs when paired with the CLDNN model is consistent with prior findings in the deep learning–based speech recognition literature [11]. The DCT step in MFCC extraction was originally introduced to decorrelate filter bank coefficients, which was essential for traditional GMM-HMM systems that assume diagonal covariance among feature dimensions. In contrast, CNN-based models do not require decorrelated input features; on the contrary, they leverage local correlations across adjacent frequency bands through their convolutional kernels, which jointly model neighboring frequency channels. Applying DCT therefore discards the very correlation structure that CNNs are designed to exploit, resulting in a loss of spectral information that is otherwise useful for learning higher-level acoustic representations. This explains why the filter bank–CLDNN combination achieves higher accuracy than the MFCC–CLDNN combination in our experiments.

Table 4: Comparison of different models using MFCC and filter bank features for speech emotion recognition. Avg. Acc.: Average Accuracy; Dim.: Feature Dimension; Time: Training and Testing Time (in seconds).

Method	Avg. Acc.	Feature	Dim.	Time (s)
CNN	0.60	MFCC	12	172.5
CNN-HMM	0.62	MFCC	12	215.6
CLDNN	0.76	MFCC	12	420.6
CNN	0.73	Filter bank	40	374.3
CNN-HMM	0.75	Filter bank	40	425.3
CLDNN	0.82	Filter bank	40	518.4

Finally, the results are visualized using a boxplot (Fig. 11), which provides an overview of the K-fold cross-validation results, including maximum, minimum, upper and lower quartiles, and outliers. Fig. 12 shows a sample boxplot used in this study.

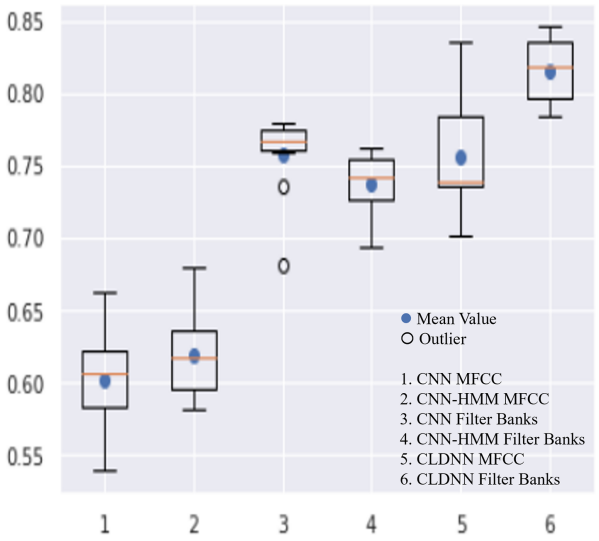

Figure 11: Boxplot representation of model performance in speech emotion recognition using K-fold cross-validation. • indicates the mean value, ○ represents outliers. Model types 1–6 are listed in the figure.

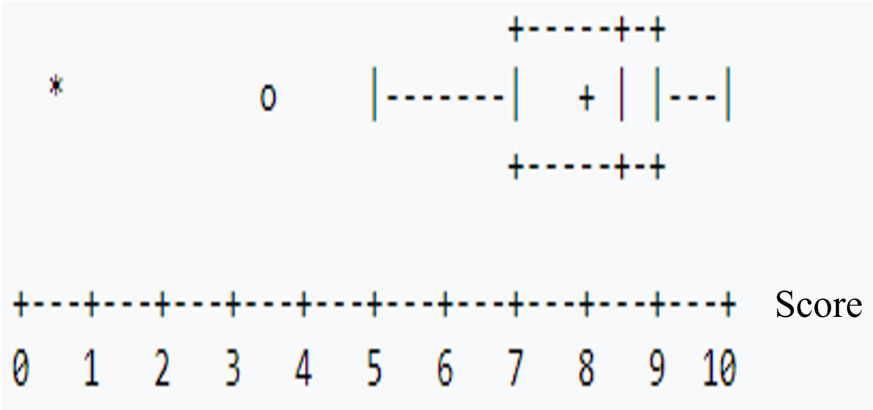

Figure 12: Boxplot example.

4.4 Continuous Facial Emotion Recognition Results

This section aims to determine whether single images or sequences of continuous frames yield better emotion recognition from video clips. For single-frame recognition, the midpoint frame, often showing the clearest emotion, is selected for model training. The CNN and InceptionV3 models are then evaluated, with single-frame recognition accuracy (without temporal modeling) reported in Table 5. For sequence-based recognition, the weights from the single-image models were saved and applied to continuous recognition. Specifically, N representative frames are extracted from each video using uniform temporal sampling: given a video of total length T frames, the i-th sampled frame is taken at index $\left[i \times \frac{T}{N}\right]$ for i = 0, 1, . . ., N − 1, ensuring that the selected frames are evenly distributed across the entire utterance from speech onset to offset. Facial regions are then extracted and preprocessed for each sampled frame, and each frame is classified using the pre-trained CNN or InceptionV3 model. The resulting sequence of frame-level emotion predictions is fed into temporal models—namely LSTM and HMM—for sequence-level emotion recognition, The corresponding results are reported in Table 6.

Table 5: Single image-based comparison (10-fold cross-validation).

Method	Average Accuracy	Number of Frames
CNN	0.61	1
InceptionV3	0.84	1

Table 6: Temporal image-based comparison (10-fold cross-validation).

Method	Average Accuracy	Number of Frames
CNN-LSTM	0.68	8
InceptionV3-LSTM	0.84	8
InceptionV3-LSTM	0.92	20
InceptionV3-HMM	0.93	40
InceptionV3-LSTM	0.94	40

The choice of N = 40 was determined by both the temporal characteristics of the RAVDESS dataset and a preliminary parameter study. Since the shortest video clip in RAVDESS is approximately 2 s long at 30 frames per second (i.e., approximately 60 frames), N must not exceed this lower bound to ensure that uniform sampling remains feasible across all clips. As shown in Table 6, recognition accuracy increases with N and saturates around N = 40, beyond which further increases yield diminishing returns. Therefore, N = 40 was selected as the optimal trade-off between recognition accuracy and computational efficiency. The best performance was ultimately achieved by combining InceptionV3 with LSTM under N = 40.

4.5 Time Series Model Comparison

This section evaluates continuous facial emotion recognition using 10-fold cross-validation, comparing accuracy and training time for 40-frame sequences over 300 iterations (Table 7). LSTM achieves higher and more stable accuracy but trains slower than HMM. HMM trains faster but shows lower and more variable accuracy, making it potentially better for real-time applications. Accuracy distribution and variability are illustrated in the box plot in Fig. 13.

Table 7: Accuracy results of each fold for InceptionV3-LSTM and inceptionV3-HMM (10-fold cross-validation).

Method	InceptionV3-LSTM	InceptionV3-HMM
Fold	**Accuracy**	**Accuracy**
Fold 1	0.945	0.942
Fold 2	0.941	0.914
Fold 3	0.951	0.947
Fold 4	0.943	0.901
Fold 5	0.939	0.938
Fold 6	0.943	0.924
Fold 7	0.945	0.942
Fold 8	0.945	0.938
Fold 9	0.943	0.934
Fold 10	0.941	0.897
Mean	0.943	0.929
SD	0.003	0.3
Variance	0.0001	0.0009
Time (s)	294.8	29.29

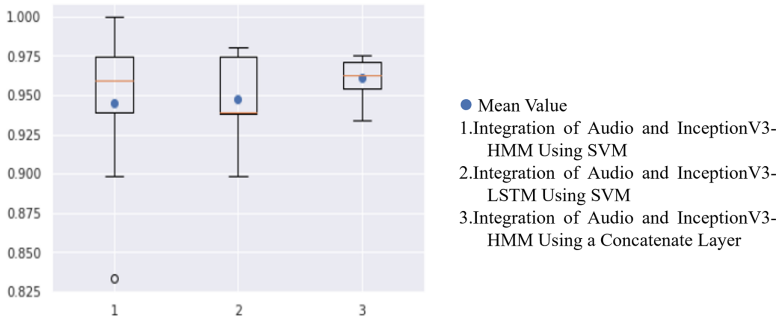

Figure 13: Box plot comparison of integrated emotion recognition models.

4.6 Comparison of Multimodal Model Integration

This section evaluates the effectiveness of integrating speech and visual recognition results to improve emotion recognition accuracy and compares different model strategies. Two fusion strategies are investigated in this study: decision-level fusion and model-level fusion. In decision-level fusion, an SVM is employed to combine the highest-performing speech model—the 40-dimensional filter banks features with CLDNN—with the two highest-performing sequential image-based models. The recognition outputs of them two modalities are stored as a two-dimensional vector (speech × video) and classified using an SVM under 10-fold cross-validation. In model-level fusion, the two modalities are integrated during model training by concatenating the fully-connected layer outputs of the speech and image branches via a Concatenate layer, followed by a final Softmax layer for classification. This strategy is introduced to mitigate the limitations of decision-level fusion, particularly the higher accuracy variability observed under cross-validation when training data are limited. The experimental results of all three configurations are summarized in Tables 8–10. Specifically, Table 8 reports the results obtained when both modalities are modeled using HMM and integrated through decision-level fusion; Table 9 reports the corresponding results when HMM is replaced

by LSTM under the same decision-level fusion strategy; and Table 10 reports the results when the LSTM-based models are integrated through model-level fusion. Comparing Tables 8 and 9 isolates the effect of the temporal modeling choice (HMM vs. LSTM), whereas comparing Tables 9 and 10 isolates the effect of the fusion strategy (decision-level vs. model-level).

Table 8: Performance comparison of speech, sequential image, and decision-level fusion emotion recognition models based on HMM, evaluated by 10-fold cross-validation on the RAVDESS dataset.

Category	**Speech**	**Sequential Image**	**Integrated Speech and Continuous Video**
Method	**CNN-HMM**	**InceptionV3-HMM**	**CNN-HMM + InceptionV3-HMM via SVM**
Fold	**Accuracy**	**Accuracy**	**Accuracy**
Fold 1	0.737	0.942	1
Fold 2	0.711	0.914	0.980
Fold 3	0.835	0.947	0.959
Fold 4	0.794	0.901	0.959
Fold 5	0.742	0.938	0.959
Fold 6	0.773	0.924	0.939
Fold 7	0.789	0.942	0.939
Fold 8	0.701	0.938	0.939
Fold 9	0.737	0.934	0.898
Fold 10	0.736	0.897	0.833
Mean	0.755	0.929	0.945
SD	0.039	0.017	0.044
Variance	0.0414	0.0003	0.0023

Table 9: Performance comparison of speech, sequential image, and decision-level fusion emotion recognition models based on LSTM, evaluated by 10-fold cross-validation on the RAVDESS dataset.

Category	**Speech**	**Sequential Image**	**Integrated Speech and Continuous Video**
Method	**CLDNN**	**InceptionV3-LSTM**	**CLDNN + InceptionV3-LSTM via SVM**
Fold	**Accuracy**	**Accuracy**	**Accuracy**
Fold 1	0.822	0.945	0.980
Fold 2	0.841	0.941	0.980
Fold 3	0.785	0.951	0.939
Fold 4	0.822	0.943	0.939
Fold 5	0.798	0.939	0.959
Fold 6	0.847	0.943	0.918
Fold 7	0.847	0.945	0.980
Fold 8	0.815	0.945	0.939
Fold 9	0.796	0.943	0.898
Fold 10	0.784	0.941	0.938
Mean	0.816	0.943	0.947
SD.	0.023	0.0031	0.026
Variance	0.0006	0.00001	0.0007

Table 10: Performance comparison of speech, sequential image, and model-level fusion emotion recognition models based on LSTM, evaluated by 10-fold cross-validation on the RAVDESS dataset.

Category	Speech	Sequential Image	Integrated Speech and Continuous Video
Method	**CLDNN**	**InceptionV3-LSTM**	**CLDNN + InceptionV3-LSTM via Concatenate layer**
Fold	**Accuracy**	**Accuracy**	**Accuracy**
Fold 1	0.822	0.980	0.971
Fold 2	0.841	0.980	0.955
Fold 3	0.785	0.939	0.971
Fold 4	0.822	0.939	0.954
Fold 5	0.798	0.959	0.954
Fold 6	0.847	0.918	0.971
Fold 7	0.847	0.980	0.975
Fold 8	0.815	0.939	0.934
Fold 9	0.796	0.898	0.963
Fold 10	0.784	0.938	0.962
Mean	0.816	0.947	0.960
SD	0.023	0.026	0.0116
Variance	0.0006	0.0008	0.00015
Time (s)	518.4	294.08	1194.95

Table 11 shows 10-fold cross-validation results, indicating that model-level fusion is more accurate and stable than decision-level fusion. Box plot visualizations in Fig. 13 further illustrate these results. The best performance is achieved by combining 40-dimensional filter bank speech features with a CLDNN model and video-based recognition models.

Table 11: 10-fold cross-validation performance comparison of integrated models.

Method	Audio + InceptionV3-HMM using SVM	Audio + InceptionV3-LSTM using SVM	Audio + InceptionV3-LSTM using concatenate layer
Fold	**Accuracy**	**Accuracy**	**Accuracy**
Fold 1	1	0.980	0.971
Fold 2	0.980	0.980	0.955
Fold 3	0.959	0.939	0.971
Fold 4	0.959	0.939	0.954
Fold 5	0.959	0.959	0.954
Fold 6	0.939	0.918	0.971
Fold 7	0.939	0.980	0.975
Fold 8	0.939	0.939	0.934
Fold 9	0.898	0.898	0.963
Fold 10	0.833	0.938	0.962
Mean	0.945	0.947	0.960
SD	0.044	0.026	0.0116
Variance	0.0023	0.0007	0.00015

As shown in Tables 8–10, the integrated multimodal models consistently outperform their unimodal counterparts across all configurations, confirming the effectiveness of fusing speech and visual modalities. A comparison between Tables 8 and 9 indicates that LSTM-based temporal modeling yields higher mean accuracy and lower variance than HMM-based modeling, suggesting LSTM's superior capability in capturing temporal dependencies in emotional expressions. Furthermore, comparing Tables 9 and 10 reveals that model-level fusion achieves higher mean accuracy (0.960 vs. 0.947) and substantially lower variance (0.00015 vs. 0.0007) than decision-level fusion, which is consistent with the observation that decision-level fusion is more sensitive to limited training data and may lose feature-level information during the SVM-based integration step.

4.7 Comparative Analysis of Previous Studies

This study compares emotion recognition methods across three categories: speech-based, vision-based, and multimodal models, against our proposed approach. For speech-based recognition, previous SVM methods show about 3% lower accuracy than ours. In vision-based recognition, reference uses CNN-LSTM, while our InceptionV3-LSTM improves accuracy by roughly 3%. For multimodal recognition, reference [31] employs HMM with speech and still images using an error-weighted classifier, and reference [7] uses CLDNN for speech and CNN for images, achieving 86% accuracy—still 10% lower than our method. Detailed comparisons are presented in Tables 12–14.

Table 12: Comparison of speech emotion recognition methods in the literature.

Study	Method	Accuracy	Year
[45] Deshmukh et al.	SVM	79%	2019
Ours	CLDNN	82%	–

Table 13: Comparison of facial emotion recognition methods in the literature.

Study	Method	Accuracy	Year
[46] Li et al.	CNN-LSTM	91%	2019
Ours	Inception V3-LSTM	94%	–

Table 14: Comparison of multimodal emotion recognition methods in the literature.

Study	Method			Accuracy	Year
	Speech	Sequential image	Integration		
[31] Liu	HMM	HMM	Error-Weighted Classifier	82%	2010
[8] Chen	CNN-LSTM	CNN	Look-up table	86%	2019
Ours	CLDNN	Inception V3-LSTM	Concatenate	96%	–

4.8 Analysis of Experimental Results

This section provides a mechanism-level analysis of the experimental results presented in Sections 4.1–4.7. While differences in datasets, frame numbers, and implementations across prior studies may limit direct comparisons, the analyses below offer interpretive insights into the design choices that drive the observed performance.

4.8.1 Effectiveness of the CLDNN Architecture for Speech Emotion Recognition

As shown in Tables 3 and 4, the CLDNN model consistently outperformed the alternative speech-only architectures evaluated in this study. This advantage can be attributed to its hybrid design: the convolutional layers extract robust local spectral patterns from the input filter bank features, the LSTM layer captures long-range temporal dependencies in the speech signal, and the fully connected layers learn discriminative high-level representations. By integrating the complementary strengths of these three architectural paradigms within a single end-to-end framework, the CLDNN achieves superior recognition performance to its individual components.

4.8.2 LSTM vs. Discrete HMM as Temporal Backbones

A central observation of this study concerns the trade-off between Discrete HMM and LSTM. As reported in Tables 8 and 9, the LSTM-based pipeline achieves higher accuracy and lower variance than its HMM counterpart under 10-fold cross-validation. This performance gap can be attributed to several factors. First, Discrete HMMs assume the Markov property—where the current state depends only on the immediately preceding state—and operate over a discrete state space, which limits their ability to model long-range and nonlinear temporal dependencies. In contrast, LSTMs employ gating mechanisms that selectively retain or discard information across arbitrarily long sequences. In the context of the present 40-frame sequences, a facial emotional expression during a 2–5 s utterance typically unfolds in three temporal phases—onset, apex, and offset—each spanning roughly 10–20 frames. The LSTM's gating mechanism enables it to retain information from the neutral baseline at the onset, accumulate the rising intensity at the apex, and integrate the entire trajectory into the final classification—dependencies that the first-order Markov assumption of the HMM cannot directly capture. Second, the discretization step required by Discrete HMMs—quantizing continuous feature vectors into a finite codebook—inevitably introduces information loss, whereas LSTMs operate directly on continuous representations. Third, HMM parameters are typically estimated via the Expectation-Maximization (EM) algorithm, which is susceptible to local optima, whereas LSTMs are trained end-to-end via backpropagation, allowing them to learn task-specific representations under sufficient training data. Furthermore, the LSTM's gating mechanism provides robustness against noisy frame-level predictions: when an individual frame is momentarily misclassified by InceptionV3 due to motion blur or partial occlusion, the LSTM can attenuate its influence based on disagreement with neighboring frames, whereas the HMM relies solely on the smoothness imposed by the transition matrix.

4.8.3 Computational Efficiency of the Discrete HMM

Despite its lower accuracy, the HMM exhibits clear advantages in computational efficiency, requiring up to ten times less time for both training and inference, and naturally accommodating variable-length input sequences. These advantages stem from the fundamentally lower computational cost of HMM inference algorithms (e.g., Viterbi and Forward–Backward), which involve matrix operations of complexity $O(N^2T)$, where N is the number of hidden states and T is the sequence length. In contrast, LSTM inference requires multiple matrix multiplications and nonlinear activations per time step, and training additionally incurs the cost of backpropagation through time. Furthermore, HMMs typically contain orders of magnitude fewer parameters than LSTMs, contributing to their faster convergence and lower memory footprint. Their generative formulation also enables operation on sequences of arbitrary length without requiring fixed-size input windows or padding. These observations suggest a clear application-oriented decision boundary: LSTMs are preferred for tasks requiring high accuracy and robustness, whereas HMMs remain a practical alternative for real-time or resource-constrained applications such as embedded systems or mobile deployment.

4.8.4 Effectiveness of Multimodal Fusion

The experimental results consistently demonstrate that fusion models combining speech and sequential image data achieve significantly higher recognition rates than either modality alone. This improvement is consistent with the complementary nature of the two modalities: speech conveys prosodic and tonal cues, while facial expressions convey visual cues such as muscle movements and micro-expressions. Each modality captures partially independent aspects of emotional expression, and their combination provides a more complete representation that is also more robust to modality-specific noise. Among the two fusion strategies investigated, model-level fusion via a concatenation layer—merging outputs from the speech CLDNN and the sequential-image InceptionV3–LSTM (40 frames)—achieved the highest mean accuracy and lowest variance, outperforming decision-level SVM fusion. This advantage likely stems from the preservation of richer feature-level information, which decision-level fusion would otherwise discard after individual classification. As shown in Table 14, this configuration achieves the highest reported accuracy among the studies surveyed, providing a useful reference point for future research in multimodal emotion recognition.

5 Conclusions

This study proposed a multimodal emotion recognition framework that integrates speech and sequential facial image data through two complementary fusion strategies. Experimental results on the RAVDESS dataset demonstrate that the proposed framework achieves substantially higher recognition accuracy (96%) than its single-image counterpart (91%), confirming the effectiveness of multi-frame sequential modeling and multimodal fusion. As discussed in Section 4, model-level fusion via concatenation outperforms decision-level SVM fusion, and LSTM provides higher accuracy than Discrete HMM at the cost of greater computational overhead—suggesting an application-oriented decision boundary where LSTMs are preferred for accuracy-critical tasks and HMMs remain practical for resource-constrained deployments.

Limitations. All experiments were conducted on the RAVDESS dataset, which was recorded in a controlled studio environment with clean audio, uniform illumination, and acted emotional expressions. Recognition accuracy under real-world conditions—involving background noise, varying illumination, and low-resolution images—remains unverified and may be substantially lower.

Future Work. Future research will (i) evaluate the proposed framework on naturalistic in-the-wild datasets such as IEMOCAP and AffectNet, (ii) incorporate noise-, illumination-, and resolution-based data augmentation strategies during training, and (iii) explore the integration of transformer-based architectures and self-supervised pre-trained encoders (e.g., Wav2Vec 2.0, HuBERT, ViT) to further enhance recognition robustness and reduce reliance on hand-crafted features.

Acknowledgement: This research work was supported by the Ministry of Science and Technology of the Republic of China under contract NSTC 114-2221-E-390-005-.

Funding Statement: This research work was supported by the Ministry of Science and Technology of the Republic of China under contract NSTC 114-2221-E-390-005-.

Author Contributions: Conceptualization, methodology, formal analysis, writing—review and editing, Shing-Tai Pan; software, validation, data curation, writing—original draft preparation, Yi-Zhen Huang and Zhi-Qing Chen. All authors reviewed and approved the final version of the manuscript.

Availability of Data and Materials: The data that support the findings of this study are openly available in Kaggle repository at https://www.kaggle.com/datasets/orvile/ravdess-dataset.

Ethics Approval: Not applicable.

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

References

1. Ekman P, Friesen WV. Facial action coding system (FACS). Palo Alto, CA, USA: Consulting Psychologists Press; 1978.
2. Mehrabian A, Ferris SR. Inference of attitudes from nonverbal communication in two channels. J Consult Psychol. 1967;31(3):248–52. doi:10.1037/h0024648.
3. Noroozi F, Marjanovic M, Njegus A, Escalera S, Anbarjafari G. Audio-visual emotion recognition in video clips. IEEE Trans Affect Comput. 2019;10(1):60–75. doi:10.1109/TAFFC.2017.2713783.
4. Ayvaz U, Gürüler H, Khan F, Ahmed N, Whangbo T, Akmalbek Bobomirzaevich A. Automatic speaker recognition using mel-frequency cepstral coefficients through machine learning. Comput Mater Contin. 2022;71(3):5511–21. doi:10.32604/cmc.2022.023278.
5. Nisar S, Asghar Khan M, Algarni F, Wakeel A, Uddin MI, Ullah I. Speech recognition-based automated visual acuity testing with adaptive mel filter bank. Comput Mater Contin. 2022;70(2):2991–3004. doi:10.32604/cmc.2022.020376.
6. Zhou Y, Liu B, Liu Y, Jiao J. Filter bank networks for few-shot class-incremental learning. Comput Model Eng Sci. 2023;137(1):647–68. doi:10.32604/cmes.2023.026745.
7. Gemello R, Albesano D, Mana F. Multi-source neural networks for speech recognition. In: Proceedings of the IJCNN'99. International Joint Conference on Neural Networks; 1999 Jul 10–16; Washington, DC, USA. p. 2946–9. doi:10.1109/IJCNN.1999.835942.
8. Chen LC. Integration of image and speech for emotion recognition based on deep learning [master's thesis]. Kaohsiung, Taiwan: National University of Kaohsiung; 2019.
9. Graves A, Jaitly N, Mohamed AR. Hybrid speech recognition with deep bidirectional LSTM. In: Proceedings of the 2013 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU 2013); 2013 Dec 8–12; Olomouc, Czech Republic. p. 273–8.
10. Chen FR, Wilcox LD, Bloomberg DS. A comparison of discrete and continuous hidden Markov models for phrase spotting in text images. In: Proceedings of 3rd International Conference on Document Analysis and Recognition; 1995 Aug 14–16; Montreal, QC, Canada. p. 398–402. doi:10.1109/ICDAR.1995.599022.
11. Sainath TN, Vinyals O, Senior A, Sak H. Convolutional, long short-term memory, fully connected deep neural networks. IEEE ICASSP. 2015;83(2):166–79. doi:10.1109/icassp.2015.7178838.
12. Lin YL, Wei G. Speech emotion recognition based on HMM and SVM. In: Proceedings of the 2005 International Conference on Machine Learning and Cybernetics; 2005 Aug 18–21; Guangzhou, China. p. 4898–901. doi:10.1109/ICMLC.2005.1527805.
13. Livingstone SR, Russo FA. The ryerson audio-visual database of emotional speech and song (RAVDESS): a dynamic, multimodal set of facial and vocal expressions in North American English. PLoS One. 2018;13(5):e0196391. doi:10.32920/25412950.v1.
14. Fahad MS, Deepak A, Pradhan G, Yadav J. DNN-HMM-based speaker-adaptive emotion recognition using MFCC and epoch-based features. Circuits Syst Signal Process. 2021;40(1):466–89. doi:10.1007/s00034-020-01486-8.
15. Yegnanarayana B, Gangashetty SV. Epoch-based analysis of speech signals. Sadhana. 2011;36(5):651–97. doi:10.1007/s12046-011-0046-0.
16. Wu X, Liu S, Cao Y, Li X, Yu J, Dai D, et al. Speech emotion recognition using capsule networks. In: Proceedings of the ICASSP 2019–2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP); 2019 May 12–17; Brighton, UK. p. 6695–9.
17. Xiang C, Zhang L, Tang Y, Zou W, Xu C. MS-CapsNet: a novel multi-scale capsule network. IEEE Signal Process Lett. 2018;25(12):1850–4. doi:10.1109/lsp.2018.2873892.
18. Liu J, Liu Z, Wang L, Guo L, Dang J. Speech emotion recognition with local-global aware deep representation learning. In: Proceedings of the ICASSP 2020—2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP); 2020 May 4–8; Barcelona, Spain. p. 7174–8. doi:10.1109/icassp40776.2020.9053192.

19. Pang C, Liu H, Li X. Multitask learning of time-frequency CNN for sound source localization. IEEE Access. 2019;7:40725–37. doi:10.1109/ACCESS.2019.2905617.
20. Jafarzadeh P, Rostami AM, Choobdar P. Speaker emotion recognition: leveraging self-supervised models for feature extraction using Wav2Vec2 and HuBERT. arXiv:2411.02964. 2024.
21. Dabbabi K, Mars A. Self-supervised learning for speech emotion recognition task using audio-visual features and distil hubert model on BAVED and RAVDESS databases. J Syst Sci Syst Eng. 2024;33(5):576–606. doi:10.1007/s11518-024-5607-y.
22. Ma X, Lin W, Huang D, Dong M, Li H. Facial emotion recognition. In: Proceedings of the 2017 IEEE 2nd International Conference on Signal and Image Processing (ICSIP); 2017 Aug 4–6; Singapore. p. 77–81. doi:10.1109/SIPROCESS.2017.8124509.
23. Li K, Wang X, Wang H. Hierarchical optimization method for federated learning with feature alignment and decision fusion. Comput Mater Contin. 2024;81(1):1391–407. doi:10.32604/cmc.2024.054484.
24. Zhang H, Jolfaei A, Alazab M. A face emotion recognition method using convolutional neural network and image edge computing. IEEE Access. 2019;7:159081–9. doi:10.1109/ACCESS.2019.2949741.
25. Javed Mehedi Shamrat FM, Majumder A, Antu PR, Barmon SK, Nowrin I, Ranjan R. Human face recognition applying haar cascade classifier. In: Pervasive computing and social networking. Singapore: Springer Nature; 2022. p. 143–57. doi:10.1007/978-981-16-5640-8_12.
26. Wang T, Huang J, Zhang H, Sun Q. Visual commonsense R-CNN. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2020 Jun 13–19; Seattle, WA, USA; 2020. p. 10757–67. doi:10.1109/cvpr42600.2020.01077.
27. Girshick R. Fast R-CNN. In: Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV); 2015 Dec 7–13; Santiago, Chile. p. 1440–8. doi:10.1109/ICCV.2015.169.
28. Purakkadavath S, Chacko T. Human-computer interaction through hybrid inception V3 and LSTM based facial emotion recognition: an experimental study. Premier J Sci. 2025;15:100213. doi:10.70389/pjs.100213.
29. Tian Y, Zhu J, Yao H, Chen D. Facial expression recognition based on vision transformer with hybrid local attention. Appl Sci. 2024;14(15):6471. doi:10.3390/app14156471.
30. Kus I, Kocak C, Keles A. A systematic review of vision transformer and explainable AI advances in multimodal facial expression recognition. Intell Syst Appl. 2026;29(1):200615. doi:10.1016/j.iswa.2025.200615.
31. Liu CJ. Audio-visual emotion recognition based on hidden markov model and error-weighted classifier combination [master's thesis]. Tainan, Taiwan: National Cheng Kung University; 2010.
32. Villarreal M, Lee MD. A Coupled Hidden Markov Model framework for measuring the dynamics of categorization. J Math Psychol. 2024;123(3):102884. doi:10.1016/j.jmp.2024.102884.
33. Othman H, Aboulnasr T. A semi-continuous state-transition probability HMM-based voice activity detector. EURASIP J Audio Speech Music Process. 2007;2007(1):043218–7. doi:10.1155/2007/43218.
34. Ghaleb E, Popa M, Asteriadis S. Metric learning-based multimodal audio-visual emotion recognition. IEEE Multimed. 2020;27(1):37–48. doi:10.1109/MMUL.2019.2960219.
35. Sánchez J, Perronnin F, Mensink T, Verbeek J. Image classification with the fisher vector: theory and practice. Int J Comput Vis. 2013;105(3):222–45. doi:10.1007/s11263-013-0636-x.
36. Yi MH, Kwak KC, Shin JH. HyFusER: hybrid multimodal transformer for emotion recognition using dual cross modal attention. Appl Sci. 2025;15(3):1053. doi:10.3390/app15031053.
37. Ramaswamy MPA, Palaniswamy S. Multimodal emotion recognition: a comprehensive review, trends, and challenges. Wiley Interdiscip Rev Data Min Knowl Discov. 2024;14(6):e1563. doi:10.1002/widm.1563.
38. Lian H, Lu C, Li S, Zhao Y, Tang C, Zong Y. A survey of deep learning-based multimodal emotion recognition: speech, text, and face. Entropy. 2023;25(10):1440. doi:10.3390/e25101440.
39. Szegedy C, Vanhoucke V, Ioffe S, Shlens J, Wojna Z. Rethinking the inception architecture for computer vision. In: Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2016 Jun 27–30; Las Vegas, NV, USA. p. 2818–26. doi:10.1109/CVPR.2016.308.

40. Chen J, Yuan C, Cui C, Xia Z, Sun X, Akilan T. A lightweight convolutional neural network with representation self-challenge for fingerprint liveness detection. Comput Mater Continua. 2022;73(1):719–33. doi:10.32604/cmc.2022.027984.
41. Mudaliar NK, Hegde K, Ramesh A, Patil V. Visual speech recognition: a deep learning approach. In: Proceedings of the 2020 5th International Conference on Communication and Electronics Systems (ICCES); 2020 Jun 10–12; Coimbatore, India. p. 1218–21. doi:10.1109/icces48766.2020.9137926.
42. Nagi J, Ducatelle F, Di Caro GA, Cireşan D, Meier U, Giusti A, et al. Max-pooling convolutional neural networks for vision-based hand gesture recognition. In: Proceedings of the 2011 IEEE International Conference on Signal and Image Processing Applications (ICSIPA); 2011 Nov 16–18; Kuala Lumpur, Malaysia. p. 342–7. doi:10.1109/ICSIPA.2011.6144164.
43. Manalu HV, Rifai AP. Detection of human emotions through facial expressions using hybrid convolutional neural network-recurrent neural network algorithm. Intell Syst Appl. 2024;21(1036):200339. doi:10.1016/j.iswa.2024.200339.
44. Tseng WS. FPGA-based chip with dual core for speeding up calculation of EMD: an example for speech recognition [master's thesis]. Kaohsiung, Taiwan: National Kaohsiung University; 2016.
45. Deshmukh G, Gaonkar A, Golwalkar G, Kulkarni S. Speech based emotion recognition using machine learning. In: Proceedings of the 2019 3rd International Conference on Computing Methodologies and Communication (ICCMC); 2019 Mar 27–29; Erode, India. p. 812–7. doi:10.1109/iccmc.2019.8819858.
46. Li TS, Kuo PH, Tsai TN, Luan PC. CNN and LSTM based facial expression analysis model for a humanoid robot. IEEE Access. 2019;7:93998–4011. doi:10.1109/access.2019.2928364.