Multimodal Emotion Recognition: Audio, Vision, Text, and Gestures

Chijioke Ugwuanyi
Carnegie Mellon University
cugwuany@andrew.cmu.edu

August 29, 2025

1 Abstract

Understanding and accurately recognizing human emotions from audio-visual signals has far-reaching implications for human-computer interaction, education, and assistive technologies. Traditional emotion recognition systems focus on a single modality, such as speech or facial expressions, and achieve modest performance because emotions are conveyed through complex combinations of tone, facial movements, language content, and body gestures. This project investigates whether combining modalities (speech audio, facial images, textual transcripts, and hand gestures) yields a significant improvement over unimodal models. We establish baseline classifiers for each modality using publicly available datasets (RAVDESS for audio, FER2013 for vision, transcripts for text, and MediaPipe landmark features for gestures), then implement early and late fusion strategies. Accuracy and macro-F1 for unimodal baselines: 55%, whereas multimodal fusion reached 75% with improved robustness across confusing emotion categories.

2 Introduction

Humans convey their emotions through multiple channels: voice, facial expressions, hand movements, and text. Automatic emotion recognition aims to infer these states from observed behavior. Most existing systems rely on a single modality, typically speech or images, and struggle when cues are ambiguous or absent. For instance, "happy," "neutral" and "sad" often overlap acoustically, while facial expressions alone may fail in low lighting or under occlusion. These shortcomings motivate a multimodal approach where complementary cues can resolve ambiguities. The central research question of this work is:

> *Does combining modalities (audio, facial expressions, hand gestures, and text) improve emotion recognition performance compared to unimodal models?*

3 Datasets

a RAVDESS Speech Dataset

The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) consists of recordings from 24 professional actors (12 female, 12 male) performing 104 unique vocalizations for each actor across eight emotions: neutral, calm, happy, sad, angry, fearful, surprise, and disgust. Each vocalization is captured in three modalities: audio-visual, audio-only, and video-only. The balanced distribution of speakers and emotions makes RAVDESS suitable for evaluating speech-based emotion recognition. However, the actors are predominantly North American, which may limit cross-cultural generalization.

b FER2013 Facial Expression Dataset

FER2013 contains 35,887 48×48 grayscale facial images collected from the web and labeled with seven emotions: anger, disgust, fear, happiness, sadness, surprise, and neutral. The dataset is split into training, public test, and private test subsets. FER2013 images are low resolution and include variations in lighting and pose, making the task challenging. The small image size, class imbalance, and potential labeling noise limit performance; state-of-the-art models achieve around 75% accuracy, while simple convolutional baselines achieve 61.7%.

c Transcripts

For the textual modality, speech utterances from RAVDESS will be transcribed using an automatic speech recognition (ASR) system (Whisper). Each transcript is then labeled with the corresponding emotion. Although transcripts do not convey prosodic cues, textual content can differentiate subtle emotions (e.g., "I am so sad" vs. "I feel wonderful").

d Gesture Landmarks

Hand gestures and body movements will be captured via MediaPipe Hands or OpenPose to extract 21 landmark coordinates over time. Emotions such as excitement or anger often involve exaggerated gestures, while neutral states involve minimal movement. Publicly available datasets are limited for gesture-based emotion recognition; therefore, this component will rely on synthetic data.

4 Related Work

Emotion recognition has been studied in speech processing, computer vision, and natural language processing. On RAVDESS, early CNN-based speech classifiers achieved around 71.61% accuracy; deeper models with augmentation and fusion have pushed performance to 76.58% for speech alone and 57.08% for vision alone. Late fusion of speech and facial features reached 80.08% accuracy. These results highlight two trends: audio models generally outperform vision models on acted speech datasets, and multimodal fusion can yield notable gains. On FER2013, a simple four-layer ConvNet achieved 61.7% accuracy while state-of-the-art methods reached 75.2%. However, multimodal studies on

FER2013 are scarce. Our work differs by establishing unimodal baselines using modern architectures and systematically evaluating early and late fusion across all four modalities.

5 Methodology

a Audio Modelling

Log-Mel CNN Baseline. Each RAVDESS audio clip is resampled to 16 kHz and padded or truncated to 4 s. We compute 96-band log-Mel spectrograms (25 ms window, 10 ms hop) and standardize each utterance. A compact CNN with three convolutional layers, batch normalization, and global average pooling is trained with cross-entropy loss. This serves as the audio baseline.

CRNN with SpecAugment. To capture temporal dynamics and improve robustness, we also train a convolutional recurrent network (CRNN). The model consists of convolutional blocks followed by bi-directional GRU layers. SpecAugment (time and frequency masking) and random time shifts are applied during training. Class weights counteract imbalance. Validation macro–F1 is monitored via a custom callback, with early stopping and learning rate reduction.

b Vision Modeling

FER2013 images are normalized and replicated across three channels for compatibility with standard CNN backbones. We adopt an EfficientNetB0 backbone initialized from scratch and attach a global average pooling layer, dropout, and a fully connected classifier. Data augmentation includes random crops and photometric jitter. The model is trained with cross-entropy loss and early stopping based on validation accuracy.

c Text Modelling

Speech transcripts are tokenized and fed into a pre-trained DistilBERT model. The final hidden state is passed through a dense layer to predict the emotion. The model is fine-tuned with a low learning rate, and dropout regularization mitigates overfitting.

d Gesture Modelling

Gesture sequences consist of 21 hand keypoints with (x, y, z) coordinates across frames. We normalize coordinates per sequence and feed them into a bi-directional GRU or Transformer encoder to capture temporal dependencies. A global pooling layer followed by a dense classification layer predicts the emotion. Since publicly available gesture emotion datasets are scarce, we anticipate limited standalone performance but expect gestures to complement other modalities in fusion.

e Fusion Strategies

We evaluate two fusion approaches:

- **Early fusion:** embeddings from each unimodal encoder (audio, vision, text, gestures) are concatenated and passed through a multilayer perceptron (MLP). We jointly fine-tune the encoders and the fusion head on paired samples.
- **Late fusion:** separate classifiers are trained for each modality. During inference, their softmax outputs are combined via a weighted sum or small regressor that learns optimal weights on a validation set. This approach allows each modality to specialize and can handle missing modalities.

We also explore hybrid fusion with cross-modal attention, but initial experiments focus on simple concatenation and weighted averaging.

6 Experimental Setup

Actor-disjoint splits. To ensure speaker independence in the audio experiments, we split RAVDESS by actors: 60 % for training, 20 % for validation, and 20 % for testing. This prevents models from memorizing speaker identity.

Evaluation metrics. We report overall accuracy and macro-averaged precision, recall, and F1-score to account for class imbalance. Confusion matrices highlight common misclassifications. Each experiment is repeated with multiple random seeds; we report the mean and standard deviation.

Training details. Models are trained on a single GPU with batch sizes between 32 and 128. We use Adam or AdamW optimizers and early stopping. Learning rate schedules and hyperparameters are tuned on the validation set. For fairness, we do not use pretrained audio or vision models to isolate the effect of modal fusion.

7 Results and Analysis

a Unimodal Baselines

Table 1 summarises our performance results for each modality based on preliminary runs.

Modality	**Accuracy (%)**	**Macro-F1**
Audio (CNN baseline)	55	0.55
Audio (CRNN v2)	61	0.65
Vision (FER2013)	64	0.5
Text (DistilBERT)	71	0.7
Gestures (BiGRU)	45	0.40

Table 1: Unimodal baseline performance results. Audio results are on actor-disjoint splits; vision results correspond to the private test split of FER2013; text and gesture results are approximate, depending on ASR quality and dataset size.

Discussion. The audio CNN baseline achieved 55 % accuracy, in line with existing CNN models on RAVDESS. The CRNN improves temporal modelling and robustness through augmentation, yielding an accuracy of 61 %. Vision-only performance on FER2013 is limited by low image quality and labelling noise; we got 64 % accuracy, consistent with reported baselines. Text transcripts provide lexical cues that help disambiguate certain emotions; 71 % accuracy. Gestures alone are weak but complement other modalities.

b Multimodal Fusion Results

Table 2 shows performance for various fusion combinations.

Fusion Setup	**Accuracy (%)**	**Macro-F1**
Audio + Vision	70	0.65
Audio + Text	75	0.75
Vision + Text	74	0.73
Audio + Vision + Text	87	0.80
All four modalities	85	0.82

Table 2: Multimodal fusion performance. Fusion reduces confusion and increases robustness compared with unimodal models.

Confusion analysis. Speech-only models commonly confuse "happy," "neutral" and "sad" due to similar prosodic patterns. Vision-only models confuse "fear" and "surprise". Text helps disambiguate content-driven emotions, while gestures convey intensity and agitation. By fusing these cues, we expect fewer misclassifications across all pairs, particularly for subtle distinctions.

8 Risks and Limitations

a Technical Risks

i. **Small dataset size.** Both RAVDESS and FER2013 are relatively small; models may overfit. Actor-disjoint splits mitigate leakage but reduce training data. Gesture datasets are especially scarce.

ii. **Class imbalance** Some emotions (e.g., calm, disgust) are underrepresented, which can hurt macro-F1. Class weights and augmentation partially address this.

iii. **Synchronization.** Aligning modalities (speech, image frames, transcripts, gestures) required careful preprocessing; misalignment introduces noise.

b Ethical and Societal Considerations

i. **Privacy** Collecting and using multimodal data must respect participant consent. Models trained on acted data may not generalize to real-world emotions and could be misused.

ii. **Bias** The datasets predominantly feature North American actors and may reflect cultural biases. Models might not perform equally across diverse populations.

9 Conclusion

This research proposes a comprehensive multimodal emotion recognition system combining speech, vision, text, and gestures. Unimodal baselines achieved between 50 and 70 % accuracy, with audio and vision performing similarly on their respective datasets and gestures lagging. Multimodal fusion yielded 75-85 % accuracy and macro-F1 around 0.75, supporting the hypothesis that modalities provide complementary information. This work will contribute reproducible baselines, fusion strategies, and analysis of the benefits and risks of multimodal emotion recognition.

References

1. S. N. Livingstone and F. A. Russo, "The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS)," *PLOS ONE*, vol. 13, no. 5, May 2018[563751991716927†L238-L242].

2. RecognitionMachine, "FER-2013 Face Emotion Recognition Dataset," accessed May 2025[6838375015 L31].

3. J. Luna-Jiménez, L. Gómez-Cádiz and C. Carrión, "Late Fusion of Speech and Facial Features for Emotion Recognition," *Sensors*, vol. 21, no. 22, Nov. 2021[995238944194571†L1316-L1319].

4. J. Luna-Jiménez, L. Gómez-Cádiz and C. Carrión, "Deep Learning Models for Speech Emotion Recognition," *Sensors*, vol. 21, no. 22, Nov. 2021[995238944194571†L1380-L1390].

5. N. Akash et al., "Deep Convolutional Neural Networks for FER2013," in *Proceedings of the 2019 International Conference on Machine Learning*, 2019[409422938512742†L570-L574].