MoodFusion: Multimodal Emotion Detection using Speech and Facial Expressions

Abstract

This project, titled MoodFusion, presents a multimodal system for human emotion detection using both speech and facial expressions. The system is implemented in MATLAB, combining audio features (energy and pitch) with visual features (face detection and mouth edge density) to recognize emotions such as Happy, Sad, Angry, and Neutral. The proposed fusion mechanism improves reliability compared to unimodal approaches.

1. Introduction

Human emotion recognition has become a vital component of Human–Computer Interaction (HCI), enabling systems to respond adaptively to human moods. While existing methods often rely on either speech or facial expressions, they may be error-prone under noise or occlusion. The MoodFusion system integrates both modalities for improved robustness. The project leverages MATLAB's computational and visualization capabilities to analyze signals in real time.

2. Objectives

• Develop a multimodal system combining speech and facial cues for emotion recognition.
• Analyze audio signals for pitch and energy features.
• Apply face detection (Viola–Jones algorithm) and mouth region analysis.
• Implement a fusion-based decision system for final classification.
• Demonstrate the system with real-time input from microphone and webcam.

3. Literature Review

• Speech-based emotion detection: Relies on pitch, energy, MFCC features. Limited by background noise.
• Facial expression detection: Uses computer vision methods like Viola–Jones, CNNs. Limited by lighting and occlusion.
• Multimodal fusion approaches: Combining modalities reduces error and increases reliability.

4. Methodology

System Workflow:
1. Input Acquisition: Record speech via microphone and capture face via webcam.
2. Audio Processing: Extract waveform, compute FFT Spectrum, estimate energy and pitch.
3. Facial Processing: Detect face (Viola–Jones), extract mouth region, calculate edge density.
4. Fusion Mechanism: Combine audio and facial analysis at decision level.
5. Final Classification: Emotions - Happy, Sad, Angry, Neutral.

Tools & Technologies:
MATLAB R2021a, Signal Processing Toolbox, Computer Vision Toolbox, Functions: audiorecorder, fft, vision.CascadeObjectDetector.

5. Results & Observations

• Audio features (pitch and energy) correlate with emotional intensity.
• Face detection with Viola–Jones is accurate for frontal images.
• Fusion improves accuracy compared to audio-only or face-only methods.
• Outputs include waveform + FFT plot, energy and pitch graphs, detected face, and final emotion label.

6. Applications

• Human–Computer Interaction
• Sentiment-aware systems (chatbots, assistants)
• Surveillance & Security
• Healthcare (stress/mood monitoring)
• Education (student engagement detection)

7. Future Work

• Implement deep learning models (CNN, RNN, Transformers) for better accuracy.
• Real-time multimodal fusion with continuous webcam + mic streaming.
• Extend classification to more emotions (Fear, Surprise, Disgust, etc.).
• Deploy as a desktop/mobile application.

8. Conclusion

The MoodFusion project successfully demonstrates that combining speech and facial features improves the accuracy of emotion detection. This multimodal approach lays the foundation for more sophisticated emotion-aware systems, enabling natural and adaptive human–computer interaction.

9. References

1. MATLAB Documentation – Signal Processing & Computer Vision Toolbox.
2. Viola, P., & Jones, M. (2001). Rapid object detection using a boosted cascade of simple features.
3. Zeng, Z., et al. (2009). A survey of affect recognition methods: Audio, visual, and spontaneous expressions.