Chapter 2: Literature Review

2.1 Introduction

The rapid advancement of Artificial Intelligence (AI) and Computer Vision has opened new horizons for non-invasive health monitoring. Mental health, specifically conditions like stress, anxiety, and burnout, has become a global concern. Unlike physical ailments, mental health struggles are often silent and difficult to quantify without professional diagnosis.

This chapter provides a comprehensive review of the existing literature regarding **Affective Computing**, focusing on Emotion Recognition via Facial Expressions and Speech Analysis. It examines current methodologies, existing mental health monitoring systems, and the technological gaps that persist. The goal is to establish a theoretical foundation for our proposed system, which integrates multimodal analysis (audio-visual) with smart interventions and privacy-preserving architecture.

2.2 Background

Before analyzing specific studies, it is essential to define the core concepts and technologies that underpin this project.

2.2.1 Affective Computing

Affective computing is the study and development of systems and devices that can recognize, interpret, process, and simulate human affects. It is an interdisciplinary field spanning computer science, psychology, and cognitive science. The core premise is that machines can be taught to detect emotional states through physiological and behavioral cues.

2.2.2 Facial Emotion Recognition (FER)

The face is the most visible indicator of human emotion. Research in FER often relies on the psychological theory of **Paul Ekman**, which classifies human emotions into universal categories: Happiness, Sadness, Anger, Fear, Disgust, and Surprise. Modern FER systems utilize Deep Learning, specifically Convolutional Neural Networks (CNNs), to map facial landmarks and texture changes to these emotional states.

2.2.3 Speech Emotion Recognition (SER)

While the face reveals *what* a person feels, the voice often reveals *how intensely* they feel it. SER focuses on extracting acoustic features such as:

- **Pitch and Tone:** High pitch often correlates with stress or panic.
- **Energy/Loudness:** Low energy may indicate sadness or drowsiness.
- Speech Rate: Rapid speech can signal anxiety.

These features are processed to determine the underlying emotional arousal of the speaker without needing to record or analyze the actual words spoken.

2.2.4 Multimodal Fusion

Single-modality systems (Face-only or Voice-only) often suffer from ambiguity. For example, a person might smile out of politeness while feeling stressed. Multimodal Fusion combines data from multiple sensors (Camera + Microphone) to achieve higher accuracy and robustness in emotion detection.

2.3 Review of Relevant Work

In this section, we analyze recent studies that have attempted to solve similar problems using AI. We have categorized the related work into three main domains: Multimodal Stress Detection, Voice Analysis Applications, and Integrated Mental Health Systems.

2.3.1 Multimodal Stress and Emotion Detection

The integration of multiple modalities (face and voice) is considered the state-of-the-art approach for robust emotion detection.[1]

- **Li et al. (2021)** proposed a framework named **"MUSER" (Multimodal Stress Detection using Emotion Recognition)**. Their stud[2]y investigated the interdependence between stress and basic emotions. They utilized the MuSE dataset to train a Transformer-based model that fuses acoustic and linguistic features.
    - *Finding:* They demonstrated that using "Emotion Recognition" as an auxiliary task significantly improves the accuracy of "Stress Detection," proving that stress is deeply linked to specific emotional patterns.
    - *Limitation:* Their system focused heavily on dataset analysis and classification accuracy, lacking a real-time intervention mechanism for the user.
- **Ahmed et al. (2024)** presented a systematic review of Multimodal Emotion Recognition (MER) techniques. They highlighted that while Deep Learning models (specifically CNNs fused with LSTMs) achieve high accuracy (over 85%), most

existing systems suffer from **high computational cost**, making them difficult to deploy on personal computers for real-time monitoring without significant lag.

2.3.2 Voice Analysis for Real-Time Applications

Voice is a powerful indicator of psychological state, often revealing stress before it appears on the face.

- **Arushi et al. (2021)** developed a real-time stress detection model specifically for **public speaking training**. Their system analyzed voice features such as Pitch, Amplitude Envelope, and MFCCs (Mel-Frequency Cepstral Coefficients) to detect stress in real-time.
    - *Finding:* They found that MFCCs are the most reliable indicator of "nervousness" in human speech. Their system was integrated into a Virtual Reality (VR) game to give users feedback.
    - *Relevance:* This is directly relevant to our project, as it proves that audio features alone can trigger a feedback loop (Intervention).

2.3.3 AI-Driven Mental Health Monitoring Systems

Moving beyond simple detection to complete monitoring systems is the next frontier.

- **A recent study by researchers in 2025 (IJFMR)** explored "AI-driven mental health monitoring systems". They proposed a theoretical architecture that uses IoT devices and mobile sensors to track user well-being.
    - *Finding:* The study concluded that the biggest barrier to adoption is **Privacy and Trust**. Users are hesitant to use systems that upload their data to the cloud.
    - *Gap:* The study emphasized the need for "Edge Computing" solutions where data is processed locally on the device—a core feature of our proposed system.

2.4 Relationship between the Relevant Work and Our Own Work

While existing literature, such as **Li et al. (2021)** and **Arushi et al. (2021)**, provides robust algorithms for detecting emotions, there remains a gap in integrating these technologies into a cohesive, user-centric support system. Most current research focuses heavily on the accuracy of classification rather than the practical application for the end-user.

Our proposed system builds upon these foundations but introduces three key differentiators to address the limitations identified in the review:

1. **Transition from Passive Detection to Active Intervention:**

Most relevant works stop at the "Reporting" stage, simply logging stress levels. Our system closes the loop by implementing Smart Interventions. It does not just observe the user; it acts by triggering real-time, context-aware recommendations (e.g., breathing exercises) immediately upon detecting a negative state.

2. **Privacy-First Architecture (Edge Computing):**

A common drawback in existing mental health monitoring systems is the reliance on cloud processing, which raises significant privacy concerns regarding video and audio data. Our system distinguishes itself by performing Local Processing. All analysis of facial expressions and voice tones occurs directly on the user's device, ensuring that sensitive biometric data is never transmitted externally.

3. **Multimodal Robustness:**

Many commercial applications rely on a single modality (face-only or voice-only), which can lead to false positives. By fusing visual and audio data, our system achieves a higher level of contextual awareness. For example, it can distinguish between a user who is simply concentrating (silent, neutral face) and one who is stressed (high-pitch voice, tense expression).

Table 2.1: Comparison with Existing Solutions

Feature	**Typical Existing Research**	**Our Proposed System**
Primary Goal	**Classification and Data Logging**	**Real-time Intervention and Support**
Modality	**Single (Face OR Voice)**	**Multimodal (Face + Voice)**
Action Taken	**Passive Reporting (Graphs)**	**Smart Alerts and Exercises**
Privacy	**Cloud Processing**	**Local/On-Device Processing**

2.5 Summary

In this chapter, we reviewed the theoretical foundations of Affective Computing and analyzed key studies in the field of facial and speech emotion recognition. We examined works like **Li et al. (2021)** regarding stress detection and **Arushi et al. (2021)** regarding real-time voice analysis.

Our review concludes that while accurate detection algorithms exist, there is a significant lack of systems that combine **passive monitoring** with **immediate therapeutic aid** while respecting user privacy. Our proposed project aims to bridge this gap by creating a privacy-centric desktop application that not only detects the user's mental state but actively assists them in managing it through intelligent recommendations.

The following chapter (Chapter 3: System Analysis) will define the specific requirements of the system, analyze existing solutions in depth, and present the architectural design and UML diagrams that will guide the implementation.

References

1. **Li, Y., et al.** (2021). *"MUSER: MUltimodal Stress Detection using Emotion Recognition as an Auxiliary Task"*. Proceedings of the NAACL.
2. **Arushi, et al.** (2021). *"Real-time Stress Detection Model and Voice Analysis: An Integrated VR-based Game for Training Public Speaking Skills"*. IEEE Conference on Games (CoG).
3. **Ahmed, T., et al.** (2024). *"A Systematic Review on Multimodal Emotion Recognition: Techniques, Challenges, and Future Directions"*. IEEE Access.
4. **IJFMR** (2025). *"A study on AI driven mental health monitoring system"*. International Journal for Multidisciplinary Research.