Emotion Recognition through facial and audio analysis

Software Requirements Specification

Name	Roll no	PRN no.
Sarvesh Bhaspale	333005	22110907
Omkar Halpatrao	333021	22110657
Jayesh Jaiswal	333021	22110394

Prepared for
TY Project 1
Guide: Dr. Swati Patil
Vishwakarma Institute of Information Technology, Pune

Table of Contents

1. Introduction

1.1 Purpose

The purpose of this SRS is to define the requirements for developing an Emotion Recognition System (ERS) that integrates Speech Emotion Recognition (SER) and Facial Emotion Recognition (FER) techniques. This system aims to accurately identify emotions from speech and facial expressions, contributing to diverse applications across human-computer interaction, security, healthcare, education, and more.

1.2 Scope

The ERS will process audio inputs for SER and image/video inputs for FER, leveraging advanced machine learning models and deep learning architectures. The system will employ techniques and methodologies highlighted in various research papers, ensuring robust emotion recognition across different datasets and scenarios.

1.3 Definitions, Acronyms, and Abbreviations

- SER: Speech Emotion Recognition
- FER: Facial Emotion Recognition
- CNN: Convolutional Neural Network
- RNN: Recurrent Neural Network
- BOAW: Bag of Audio Words
- MFCC: Mel Frequency Cepstral Coefficients
- IEEE: Institute of Electrical and Electronics Engineers

2. Overall Description

2.1 Product Perspective

The Emotion Recognition System (ERS) will function as a standalone system, possessing the capability to process a wide array of inputs for both Speech Emotion Recognition (SER) and Facial Emotion Recognition (FER). Interfacing seamlessly with diverse datasets, the system will leverage machine learning models to achieve precise and accurate emotion recognition. Operating in real-time, the ERS will deliver consistent and dependable emotion analysis across various scenarios and applications.

2.2 Product Functions

Speech Emotion Recognition (SER)

The system will employ advanced methodologies from diverse datasets such as CREMA-D, RAVDESS, and SAVEE for Speech Emotion Recognition (SER). Emphasis will be placed on two primary feature extraction techniques: spectrograms and Mel-spectrograms.

1. **Spectrograms:** Utilize spectrograms to transform audio signals into visual representations that display the spectrum of frequencies of a sound as it varies with time. This technique enables the system to capture the frequency content and temporal dynamics of the speech signal, forming a basis for emotion classification.

2. **Mel-spectrograms:** Implement Mel-spectrograms which emphasize critical frequencies in the speech signal through the Mel scale, enhancing the system's ability to capture nuanced vocal features associated with emotions. This feature extraction technique aims to improve the accuracy and robustness of emotion recognition.

The SER component will integrate these methodologies to process audio inputs, extracting relevant features, and using deep learning architectures for accurate emotion classification.
Facial Emotion Recognition (FER)
The system will analyze image and video inputs to perform Facial Emotion Recognition (FER) by leveraging state-of-the-art techniques such as Deep Separation Convolutional Neural Networks (DSCNN) and other advanced deep learning architectures.

1. **DSCNN:** Utilize the DSCNN technique, which focuses on interpreting facial expressions from images or video frames. This method aims to extract intricate facial features, identifying key emotional cues and patterns within the facial data.

2. **Deep Learning Architectures:** Employ deep learning architectures specifically tailored for FER, encompassing Convolutional Neural Networks (CNNs) and potentially Recurrent Neural Networks (RNNs) for understanding temporal dynamics in facial expressions. These architectures aim to capture and interpret complex facial features, facilitating accurate emotion classification.

The FER component will process image or video inputs, extract facial features, and apply advanced deep learning methodologies to recognize and classify emotions based on facial expressions.

Deep Learning Architectures
The system will implement diverse deep learning architectures, including Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and attention mechanisms to facilitate feature extraction and emotion classification across both SER and FER modules.

1. **CNNs:** Utilize CNNs for extracting hierarchical features from audio spectrograms and facial image/video data. These architectures will enable the system to automatically learn discriminative features essential for emotion recognition.

2. **RNNs:** Incorporate RNNs to capture temporal dependencies in speech and facial expressions. These architectures are crucial for understanding sequential data and long-term dependencies in emotions expressed over time.

3. **Attention Mechanisms:** Implement attention mechanisms within the deep learning architectures to focus on crucial elements within the input data, enhancing the system's capability to identify significant features contributing to emotion recognition.

These deep learning architectures will constitute the backbone of the system, enabling robust feature extraction and accurate classification of emotions in both speech and facial inputs.

2.3 User Classes and Characteristics

Developers

- **Responsibilities:** Developers form the core group involved in building, maintaining, and updating the Emotion Recognition System (ERS). They are responsible for system architecture, implementing algorithms, integrating new methodologies, and ensuring the system's reliability and scalability.

- **Characteristics:**

    - **Technical Proficiency:** Profound knowledge of machine learning, deep learning, and signal processing techniques.

    - **Programming Skills:** Competence in programming languages (Python, MATLAB, etc.) and frameworks (TensorFlow, PyTorch, etc.) relevant to machine learning and signal processing.

    - **System Maintenance:** Ability to maintain and update the ERS with the latest advancements, ensuring optimal system performance and reliability.

End Users

- **Responsibilities:** End users interact with the ERS for emotion analysis in diverse applications such as customer service, healthcare, security, and more. They utilize the system's insights to make informed decisions and take appropriate actions based on emotion analytics.

- **Characteristics:**

    - **Varied Domain Knowledge:** Users may come from different industries such as customer service, healthcare, security, etc., each requiring emotion analysis for specific purposes within their respective domains.

- **Application Proficiency:** Ability to incorporate emotion analytics provided by the ERS into their applications or workflows to enhance decision-making processes.
- **Usage Scenarios:** Users might leverage the ERS in real-time or for batch analysis, depending on their specific application needs.

2.4 Operating Environment

The Emotion Recognition System (ERS) operates within a specified environment, necessitating hardware capabilities, compatible software, and seamless integration with databases. The following aspects define the ERS operating environment:

Hardware Requirements

- **Processing Power:** The ERS requires hardware capable of running resource-intensive machine learning models efficiently. This includes CPUs or GPUs with adequate processing power to handle computations for neural networks and deep learning architectures.
- **Memory and Storage:** Sufficient RAM and storage space are essential to accommodate large datasets, model parameters, and system resources during training and inference phases.

Software Environment

- **Operating Systems:** The ERS should be compatible with common operating systems such as Windows, Linux (Ubuntu, CentOS, etc.), or macOS, ensuring broad accessibility across different platforms.
- **Development Frameworks and Libraries:** The system requires installation and compatibility with machine learning frameworks like TensorFlow, PyTorch, Keras, or scikit-learn. Additionally, it may utilize signal processing libraries (such as LibROSA) and database management systems (e.g., MySQL, PostgreSQL) for efficient data handling.

Database Integration

- **Data Access and Management:** The ERS interfaces with databases to efficiently access, manage, and store datasets used for training, validation, and testing purposes. Compatibility with various database systems ensures seamless integration and data retrieval.

Networking Requirements (if applicable)

- **Connectivity:** For cloud-based applications or distributed systems, the ERS might need stable internet connectivity to access cloud-based databases or collaborative platforms for remote development and deployment.

Security Considerations

- **Data Security:** Implementing security protocols to protect sensitive data within the system and during data transmission is critical. Encryption methods and secure data handling practices should be adhered to, following industry standards and regulatory compliance.

3. Specific Requirements

3.1 External Interface Requirements

Inputs: Accept audio files (e.g., WAV, MP3) for SER and image/video files (JPEG, MP4) for FER.
Outputs: Provide emotion classification results, possibly accompanied by visual representations or reports.

3.2 Functional Requirements

The Emotion Recognition System (ERS) is defined by various functional requirements essential for its effective operation and performance. These requirements encompass audio and facial feature extraction, machine learning integration, and real-time processing capabilities:

Audio Feature Extraction

- **SER Methodologies:** Implement methodologies such as Mel Frequency Cepstral Coefficients (MFCCs) and spectrograms to extract essential audio features that accurately represent emotions embedded within speech inputs.

Facial Feature Extraction

- **Facial Emotion Recognition (FER):** Utilize advanced techniques like Deep Separation Convolutional Neural Networks (DSCNN) and other state-of-the-art methodologies for robust and accurate extraction of facial features, enabling precise identification of emotions from image or video inputs.

Machine Learning Integration

- **Deep Learning Architectures:** Integrate Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and attention mechanisms within the system's framework. These neural network architectures will aid in the extraction of discriminative features and enable precise emotion classification from both audio and visual inputs.

Real-time Processing

- **Efficient Real-time Processing:** Ensure the ERS's capability to swiftly and reliably recognize emotions in real-time scenarios, allowing for immediate analysis and response in applications like customer service or security monitoring.

These functional requirements form the core capabilities of the Emotion Recognition System (ERS), encompassing audio and facial feature extraction methodologies, integration of advanced machine learning architectures, and the ability to perform real-time emotion recognition across various applications and scenarios.

3.3 Non-Functional Requirements

The Emotion Recognition System (ERS) is bound by several non-functional requirements that significantly impact its effectiveness, usability, and reliability across different contexts and user interactions:

Performance

- **High Accuracy:** Achieve an emotion recognition accuracy of over 85% across diverse datasets and real-world scenarios. The system should consistently identify emotions embedded **within audio and visual inputs accurately.**

Reliability

- **Consistent Recognition:** Ensure consistent and reliable emotion recognition under various environmental conditions, maintaining accuracy and stability in different contexts.

Security

- **Data Privacy:** Implement robust measures to ensure data privacy, safeguarding against unauthorized access, manipulation, or misuse of sensitive information stored within the system.

Usability

- **User Interface:** Develop an intuitive and user-friendly interface to facilitate easy interaction and interpretation of emotion analysis results, catering to users with varying levels of technical expertise.

Maintainability

- **Modular Codebase:** Create a well-structured and modular codebase, accompanied by comprehensive documentation, to ease system maintenance, updates, and future enhancements. This will enable efficient management of the system's components and functionalities.

These non-functional requirements form the foundational aspects of the Emotion Recognition System (ERS), ensuring its performance, reliability, security, usability, and maintainability to meet user needs and expectations effectively.

3.4 System Features

- **Integration of Models:** The Emotion Recognition System (ERS) integrates both Speech Emotion Recognition (SER) and Facial Emotion Recognition (FER) models, enabling comprehensive emotion analysis by synergizing audio and facial data.

- **Real-time Processing:** Ensuring swift and efficient emotion recognition in real-time scenarios, the system processes inputs promptly, enabling quick analysis of emotions in dynamic environments.

- **Cross-validation:** The system rigorously validates and tests its performance across diverse datasets and modalities. This process ensures the system's robustness and reliability in accurately recognizing emotions, irrespective of the dataset or input modality.

4. Conclusion

This Software Requirements Specification serves as a comprehensive blueprint for the development of an Emotion Recognition System (ERS) integrating Speech and Facial Emotion Recognition techniques. By delineating essential features, functionalities, and system requirements, this document acts as a guiding framework for the design and development phases of the system.

The delineated specifications encompass user requirements, functional and non-functional aspects, system interfaces, and operating environments, providing a holistic view essential for building an efficient and reliable Emotion Recognition System. Emphasizing the utilization of advanced methodologies from diverse datasets, such as CREMA-D, RAVDESS, and SAVEE, along with cutting-edge deep learning architectures, this document aligns with the latest advancements in the field of emotion analysis.

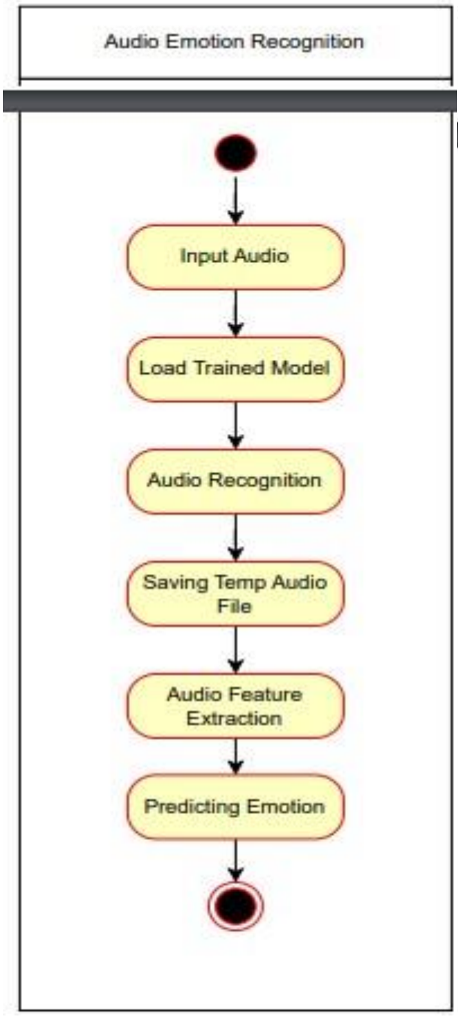
With a focus on accuracy, reliability, security, usability, and maintainability, this document lays the groundwork for the creation of an ERS capable of robustly recognizing emotions across varied datasets and real-world scenarios. The integration of advanced techniques like MFCCs, spectrograms, DSCNN, CNNs, RNNs, and attention mechanisms underscores the system's potential to accurately capture and interpret emotions from both audio and visual inputs.

In summary, this Software Requirements Specification stands as an indispensable reference, guiding the development team in creating an Emotion Recognition System that meets the highest standards of performance, usability, security, and reliability, thereby catering to diverse user needs and applications across industries.

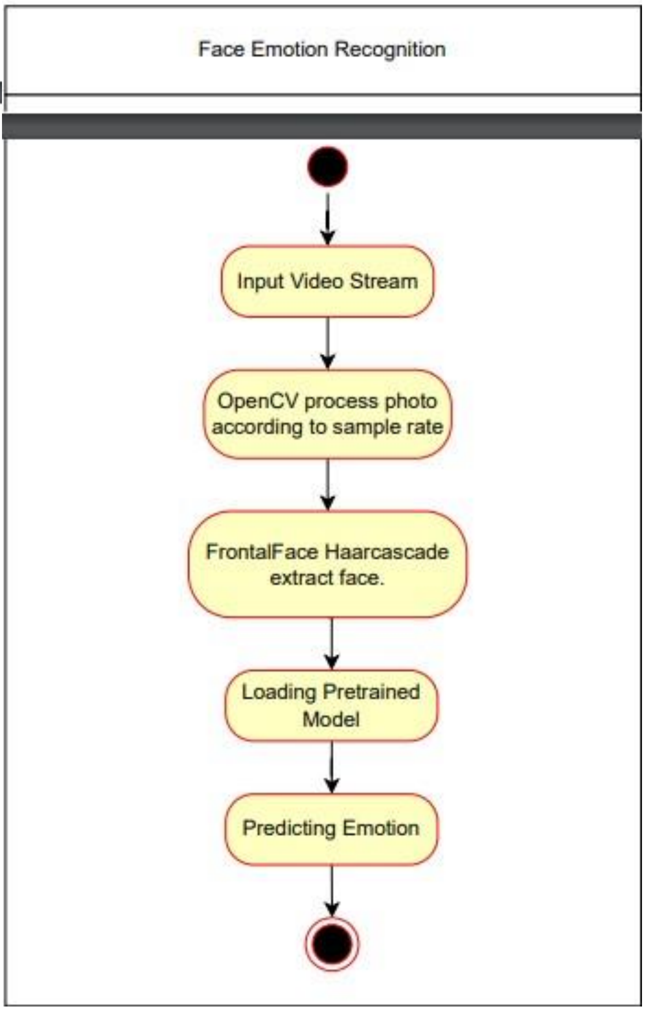
5.Appendix

A1. UML Diagrams

Activity Diagram

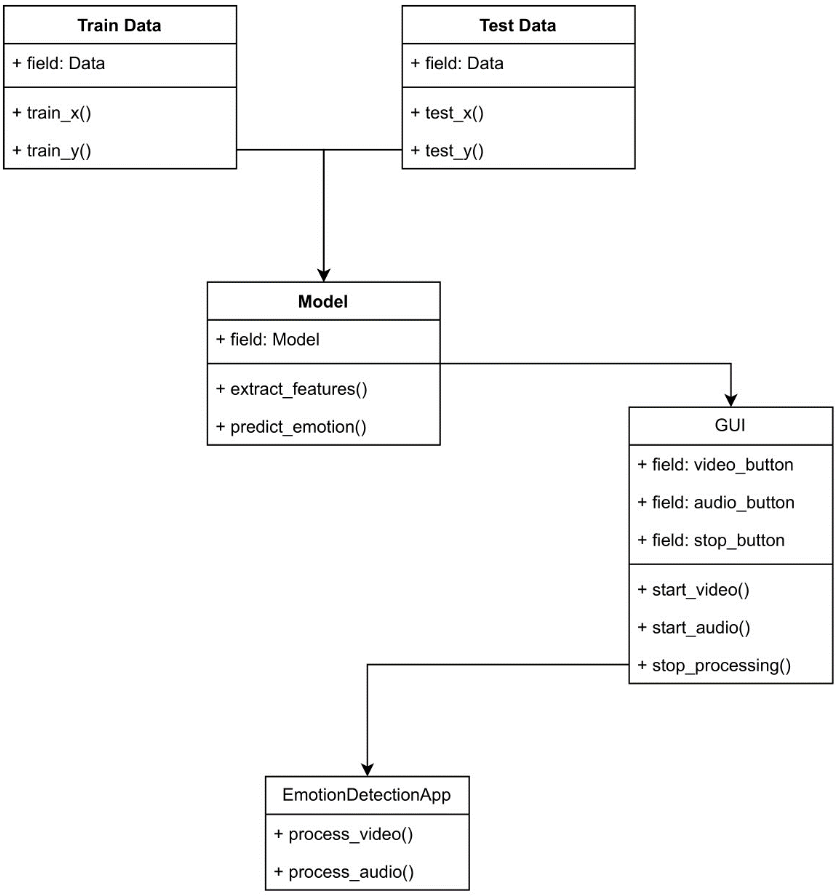


Audio

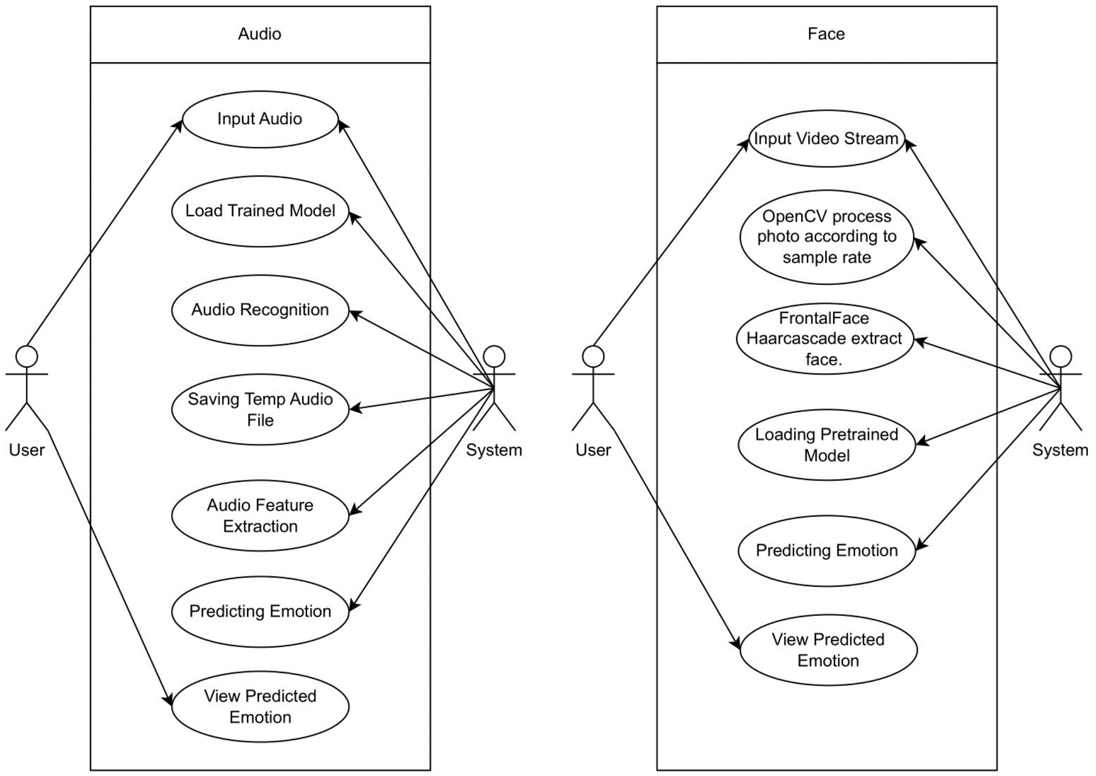


Facial

Class Diagram



Use Case Diagram



A2. References

[1] R. Abou Zafra, A. A. Abdullah, L. Alaraj, R. Albezreh, and R. Barhoum, “An experimental study in Real-time Facial Emotion Recognition on new 3RL dataset,” 2023. [Online]. Available: https://github.com/

[2] X. Wang, Y. Wang, and D. Zhang, “Complex Emotion Recognition via Facial Expressions with Label Noises Self-Cure Relation Networks,” Comput. Intell. Neurosci., vol. 2023, pp. 1–10, Jan. 2023, doi: 10.1155/2023/7850140.

[3] T. Kumar Arora et al., “Optimal Facial Feature Based Emotional Recognition Using Deep Learning Algorithm,” Comput. Intell. Neurosci., vol. 2022, 2022, doi: 10.1155/2022/8379202.

[4] D. Ammous, A. Chabbouh, A. Edhib, A. Chaari, F. Kammoun, and N. Masmoudi, “Designing an Efficient System for Emotion Recognition Using CNN,” J. Electr. Comput. Eng., vol. 2023, pp. 1–11, Sep. 2023, doi: 10.1155/2023/9351345.

[5] Z. Y. Huang et al., “A study on computer vision for facial emotion recognition,” Sci. Rep., vol. 13, no. 1, Dec. 2023, doi: 10.1038/s41598-023-35446-4.

[6] A. Pandey, A. Gupta, and R. Shyam, “FACIAL EMOTION DETECTION AND RECOGNITION,” Int. J. Eng. Appl. Sci. Technol., vol. 7, no. 1, pp. 176–179, May 2022, doi: 10.33564/IJEAST.2022.v07i01.027.

[7] M. Zielonka, A. Piastowski, A. Czyżewski, P. Nadachowski, M. Operlejn, and K. Kaczor, “Recognition of Emotions in Speech Using Convolutional Neural Networks on Different Datasets,” Electron., vol. 11, no. 22, Nov. 2022, doi: 10.3390/electronics11223831.

[8] A. S. Alluhaidan, O. Saidani, R. Jahangir, M. A. Nauman, and O. S. Neffati, “Speech Emotion Recognition through Hybrid Features and Convolutional Neural Network,” Appl. Sci., vol. 13, no. 8, Apr. 2023, doi: 10.3390/app13084750.

[9] J. L. Bautista, Y. K. Lee, and H. S. Shin, “Speech Emotion Recognition Based on Parallel CNN Attention Networks with Multi-Fold Data Augmentation,” Electron., vol. 11, no. 23, Dec. 2022, doi: 10.3390/electronics11233935.

[10] S. Chamishka et al., “A voice-based real-time emotion detection technique using recurrent neural network empowered feature modelling,” Multimed. Tools Appl., vol. 81, no. 24, pp. 35173–35194, Oct. 2022, doi: 10.1007/s11042-022-13363-4.