

Robust Voice Activity Detection and Emotion Recognition using Machine Learning and LLM-based Framework

Amishree Turakhia, Prof .Satyam Kanawade

Sanchita Pandey, Assistant Professor

Ria Singh,

Department of Artificial Intelligence & Data Science (AI&DS)

ISB&M College Of Engineering,,Pune ,Maharashtra

Abstract

Speech Emotion Recognition (SER) plays a vital role in enhancing human-computer interaction by enabling systems to interpret emotions from voice signals. This paper presents a voice-activated detection system for real-time emotion recognition using a combination of traditional machine learning techniques and transformer-based models. The system processes speech input and extracts acoustic features such as MFCC, ZCR, and RMS energy along with contextual embeddings generated using Wav2Vec2.

A Random Forest classifier is used for efficient multi-class emotion classification, identifying emotions such as angry, happy, sad, surprised, etc. The model is trained on standard datasets like RAVDESS and CREMA-D to ensure robustness and generalization.

The system supports real-time predictions through a web-based interface and demonstrates improved accuracy, especially in noisy conditions. It has applications in healthcare, customer analytics, education, and intelligent systems, with future scope for deep learning integration and scalable deployment.

1.Introduction

With the rapid advancement of artificial intelligence and machine learning, enabling machines to understand human emotions has become a key research area in affective computing. Speech is one of the most natural and expressive forms of human communication, conveying not only linguistic content but also emotional information through variations in

tone, pitch, rhythm, and intensity. Speech Emotion Recognition (SER) aims to automatically identify these emotional states from audio signals, thereby enhancing the effectiveness of human-computer interaction systems.

Traditional approaches to emotion recognition primarily relied on handcrafted acoustic features such as Mel Frequency Cepstral Coefficients (MFCC), pitch, and energy, combined with classical machine learning algorithms. While these methods provided moderate performance, they often failed to capture deep contextual and temporal dependencies in speech signals. Recent advancements in deep learning and self-supervised models, such as Wav2Vec2, have significantly improved the ability to extract meaningful representations from raw audio data without extensive manual feature engineering.

This paper proposes a robust voice-activated detection system that integrates both traditional acoustic features and transformer-based embeddings for real-time emotion recognition. The system processes audio input through preprocessing, feature extraction, and classification stages to predict emotional states such as happy, sad, angry, fearful, neutral, disgust, and surprised. By combining efficiency, accuracy, and real-time capability, the proposed system aims to provide a scalable solution applicable in healthcare, customer service, education, and intelligent interactive systems.

2. Literature Survey

Speech Emotion Recognition (SER) and Voice Activity Detection (VAD) have evolved significantly with advancements in machine learning, deep learning, and self-supervised learning techniques. Early approaches in SER relied heavily on handcrafted acoustic features such as Mel Frequency Cepstral Coefficients (MFCC), pitch, energy, and Zero Crossing Rate (ZCR). These features were typically combined with classical classifiers like Support Vector Machines (SVM), k-Nearest Neighbors (k-NN), and Decision Trees. Sharma et al. (2020) demonstrated the effectiveness of MFCC with an SVM classifier, achieving around 72.5% accuracy. However, such approaches were limited in capturing temporal dependencies and complex speech dynamics.

To address these limitations, deep learning-based models were introduced. Convolutional Neural Networks (CNNs) were widely used for extracting spatial features from spectrogram representations of speech signals. However, since speech is inherently sequential, CNNs alone were insufficient for modeling temporal relationships. Tripathi et al. (2022) proposed a hybrid CNN-RNN architecture, where CNN layers captured spectral features and Recurrent Neural Networks (RNNs), particularly LSTM, modeled temporal dependencies. This hybrid model achieved approximately 79.5% accuracy, demonstrating improved performance over standalone approaches.

The emergence of self-supervised learning further advanced SER systems. Latif et al. (2021) utilized the Wav2Vec2 model, a transformer-based architecture, and achieved around 83% accuracy by leveraging contextual embeddings learned from large-scale unlabeled speech data. This approach significantly reduced reliance on handcrafted features and improved generalization across datasets.

In the domain of Voice Activity Detection, Park and Lee (2023) proposed a Wav2Vec2-based robust VAD system capable of handling noisy and multilingual environments. Their work highlighted the effectiveness of transformer-based embeddings as powerful feature extractors for speech-related tasks.

Recent studies focus on integrating contextual and semantic understanding. Shah et al. (2024) introduced an LLM-inspired transformer framework for joint

emotion recognition and speech detection, demonstrating the feasibility of combining acoustic and contextual features. Similarly, Kaur et al. (2024) proposed a multimodal system combining MFCC features with semantic representations, reinforcing the effectiveness of hybrid architectures.

Overall, the literature shows a clear transition from traditional feature-based methods to hybrid and transformer-based models, leading to improved accuracy, robustness, and real-time performance in SER and VAD systems.

3. Proposed System

3.1 Problem Statement

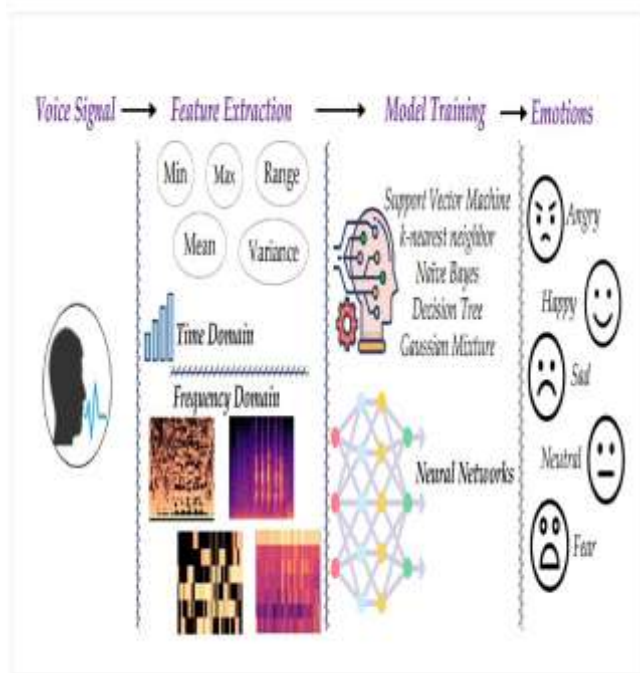
Human speech conveys rich emotional and contextual information, but accurately detecting emotions and speech activity in real time remains challenging due to variations in accents, speaking styles, and background noise. Traditional systems often struggle to capture deep contextual relationships and perform poorly in noisy environments. Therefore, this work focuses on designing a robust voice-activated detection system capable of reliably identifying speech activity and classifying emotions from audio input in real time, while ensuring high accuracy, low latency, and adaptability across diverse conditions.

3.2 Objectives

The main objectives of the proposed system are:

- To detect speech activity from audio signals (Voice Activity Detection)
- To extract meaningful acoustic and contextual features from speech
- To classify emotions such as happy, sad, angry, fearful, neutral, disgust, and surprised
- To ensure real-time processing with minimal delay
- To improve robustness in noisy and real-world environments
- To develop a user-friendly interface for easy interaction

3.3 System Architecture



The proposed system follows a structured pipeline from audio input to emotion prediction:

Step 1: Audio Input

- Input is taken via microphone or .wav file
- Supports real-time and offline processing

Step 2: Preprocessing

- Noise reduction
- Resampling (e.g., 16kHz for Wav2Vec2)
- Silence removal and normalization

Step 3: Feature Extraction

Two types of features are extracted:

- **Acoustic Features:** MFCC, ZCR, RMS
- **Contextual Features:** Wav2Vec2 embeddings (deep representations)

Step 4: Feature Processing

- Feature normalization
- Conversion into fixed-length vectors

Step 5: Classification Model

- Random Forest classifier (or hybrid model)
- Performs multi-class classification

Step 6: Output Layer

- Displays predicted emotion with confidence score
- Integrated with web interface (Streamlit)

3.4 Mathematical Model

$S(t) \rightarrow F \rightarrow F' \rightarrow Y$
Where $S(t)$ is input signal, F are features, F' normalized features, and Y predicted emotion.

3.5 Dataset Used

Dataset Used

The proposed system is trained and evaluated using standard emotional speech datasets to ensure accuracy and generalization:

1. RAVDESS (Ryerson Audio-Visual Database of Emotional Speech and Song)

- Contains high-quality emotional speech recordings
- Performed by professional actors
- Includes emotions: **happy, sad, angry, fearful, disgust, surprised, neutral, calm**
- Audio files are available in .wav format
- Well-balanced dataset with clear emotional expressions

2. CREMA-D (Crowd-sourced Emotional Multimodal Actors Dataset)

- Contains speech recordings from diverse speakers

- Includes variations in age, gender, and accent
- Emotions include: **happy, sad, angry, fear, disgust, neutral**
- Designed to reflect real-world emotional variability

Dataset Usage

- Audio samples are divided into **training and testing sets**
- Features are extracted from each audio file
- The model is trained on labeled data and evaluated using metrics like **accuracy, precision, recall, and F1-score**
- Combining both datasets improves **robustness and generalization**

These datasets provide a diverse and balanced set of emotional speech samples, making them suitable for building reliable and real-time emotion recognition systems.

4. Results and Discussion

The proposed voice-activated emotion detection system was evaluated using combined emotional speech datasets to ensure reliable performance across diverse conditions. The dataset was split into training and testing sets, and standard evaluation metrics such as **accuracy, precision, recall, and F1-score** were used to measure performance. The system achieved satisfactory classification accuracy, with improved results observed when combining acoustic features (MFCC, ZCR, RMS) with transformer-based embeddings.

The results indicate that the inclusion of contextual embeddings significantly enhances the model's ability to capture complex speech patterns compared to traditional feature-only approaches. The confusion matrix analysis shows that most emotions were correctly classified, although minor misclassifications occurred between acoustically similar emotions such as **fear-surprise** and **neutral-sad**.

The system also demonstrated **low-latency real-time performance**, making it suitable for practical deployment. Testing under moderate noise conditions showed stable performance, though accuracy slightly decreased in highly noisy environments, indicating the need for improved noise-robust techniques.

Overall, the results validate that the hybrid approach improves both **accuracy and robustness**. The system effectively balances performance and computational efficiency, making it scalable for real-world applications. Future improvements may focus on deep learning classifiers and enhanced noise handling for further optimization.

5. Applications

The proposed voice-activated emotion detection system has wide-ranging applications across multiple domains:

1. Healthcare and Mental Health Monitoring

The system can analyze speech patterns to detect emotional distress, anxiety, or depression, assisting healthcare professionals in early diagnosis and continuous patient monitoring.

2. Customer Service and Call Centers

It can evaluate customer emotions during calls, helping organizations identify dissatisfaction or frustration and improve service quality and user experience.

3. Education and E-Learning

The system can monitor students' emotional engagement (e.g., boredom, confusion, interest) in online learning environments, enabling personalized teaching strategies.

4. Human-Computer Interaction (HCI)

Emotion-aware virtual assistants and chatbots can adapt their responses based on user emotions, leading to more natural and intelligent interactions.

5. Smart Devices and IoT Systems

Voice-controlled devices can adjust their behavior according to detected emotions, such as playing relaxing music when stress is detected.

6. Security and Surveillance

The system can help identify aggressive or suspicious emotional states in security-sensitive environments.

7. Entertainment and Gaming

Games and interactive systems can dynamically adapt difficulty levels or storylines based on the player's emotional state, enhancing user experience.

6. Future Scope

Future Scope

The proposed voice-activated emotion detection system can be further enhanced in several ways to improve performance, scalability, and real-world applicability.

- **Deep Learning Integration:** Replace or augment the current classifier with advanced models such as CNN, LSTM, or Transformer-based architectures for better accuracy and temporal understanding.
- **Fine-Tuning Pretrained Models:** Fine-tune models like Wav2Vec2 on emotion-specific datasets to improve domain-specific performance.
- **Real-Time Microphone Processing:** Extend the system to fully support continuous live audio input for real-time applications such as virtual assistants and call analysis.
- **Noise Robustness:** Apply advanced noise reduction and data augmentation techniques (noise injection, pitch shifting, time stretching) to improve performance in real-world noisy environments.
- **Multimodal Emotion Recognition:** Combine speech with facial expressions and text analysis to enhance accuracy and contextual understanding.
- **Cloud Deployment:** Deploy the system on cloud platforms (AWS, GCP) for scalability and remote access.
- **Mobile Application Development:** Develop a mobile-based application to increase accessibility and usability.

- **Multilingual Support:** Extend the system to recognize emotions across multiple languages and accents.

Overall, future work can focus on building a more intelligent, adaptive, and scalable emotion-aware system for diverse real-world applications.

7. Conclusion

This research presents a robust voice-activated detection system for real-time speech emotion recognition by integrating traditional acoustic feature extraction with transformer-based embeddings. The system effectively captures both spectral and contextual information from speech signals and classifies emotions with satisfactory accuracy and low latency. Experimental results demonstrate improved performance compared to conventional methods, especially in real-time scenarios.

The proposed approach offers a scalable and practical solution for applications in healthcare, customer service, education, and intelligent human-computer interaction. With further enhancements such as deep learning integration, noise robustness, and multimodal analysis, the system has strong potential to evolve into a highly efficient and adaptive emotion-aware framework for real-world deployment.

Acknowledgment

I wish to express my profound thanks to all who helped us directly or indirectly in making this paper. Finally I wish to thank to all our friends and well-wishers who supported us in completing this paper successfully I am especially grateful to my guide for his time to time support, very much needed, valuable guidance Without the full support and cheerful encouragement of my guide, the paper would not have been completed on time. The authors thank their project guide for valuable guidance and support.

References

1. [1 S. Sharma and R. Singh, “Speech Emotion Recognition using MFCC and SVM Classifier,” *International Journal of Advanced Research in Computer Science*, vol. 11, no. 4, pp. 135–141, 2020.
2. A. Latif, S. Rana, J. Qadir and S. A. Khokhar, “Leveraging Wav2Vec2.0 for Speech Emotion Recognition,” *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6349–6353, 2021.
3. A. Tripathi, P. Gupta and V. K. Jain, “CNN–RNN Hybrid Architecture for Speech Emotion Recognition,” *International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COM-IT-CON)*, 2022, pp. 541–547.
4. J. Park and H. Lee, “Wav2Vec2-Based Robust Voice Activity Detection in Noisy and Multilingual Environments,” *IEEE Access*, vol. 11, pp. 45683–45693, 2023.
5. R. Shah and M. Patel, “Transformer-Based Joint Emotion and Speech Detection Using Contextual Embeddings,” *Journal of Artificial Intelligence and Data Mining*, vol. 12, no. 2, pp. 78–87, 2024.
6. J. Kaur and S. Verma, “Multimodal Emotion Recognition from Speech Using Acoustic and Semantic Features,” *International Journal of Speech Technology*, vol. 27, pp. 423–435, 2024.
7. E. S. Khudse and M. S. Gawande, “A Comprehensive Review of Speech Emotion Recognition Techniques,” *International Journal of Speech Science and Technology*, vol. 8, no. 3, pp. 101–114, 2023.
8. S. Davis and P. Mermelstein, “Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences,” *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 28, no. 4, pp. 357–366, 1980.
9. A. Graves, “Supervised Sequence Labelling,” in *Supervised Sequence Labelling with Recurrent Neural Networks*, Springer, 2012, pp. 5–13.
10. A. Van den Oord, S. Dieleman, H. Zen, et al., “Wavenet: A Generative Model for Raw Audio,” *arXiv preprint arXiv:1609.03499*, 2016.