International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 09 Issue: 08 | Aug - 2025 SJIF Rating: 8.586 ISSN: 2582-3930

Voice Mood Detector

Harshavardhan B L [1], Prof. Rajeshwari N [2]

[1] Student, Department of MCA, Bangalore Institute of Technology, Karnataka, India

[2] Professor, Department of MCA, Bangalore Institute of Technology, Karnataka, India

ABSTRACT

The Voice Mood Detector is a web-based system designed to analyze speech and recognize human emotions using machine learning and signal processing. Unlike traditional speech-to-text systems, it focuses on tone, pitch, rhythm, and energy to classify emotions such as happy, sad, angry, calm, fearful, and disgust. The system integrates a React and TypeScript frontend with a Flask backend, offering real-time audio processing through feature extraction methods like MFCC, Chroma, and Mel-spectrogram. A Multi-Layer Perceptron (MLP) classifier, trained on the RAVDESS dataset, achieves an accuracy of over 85%, ensuring reliable results. Key features include drag-and-drop audio upload, secure authentication, real-time prediction, confidence scoring, and history tracking. The system prioritizes privacy by processing data locally rather than relying on cloud services. With applications in mental health monitoring, customer service, and education, the project demonstrates how artificial intelligence can enhance human–computer interaction and pave the way for advanced emotion-aware technologies.

1. INTRODUCTION

Human communication is not limited to words alone; emotions play a vital role in expressing thoughts, feelings, and intentions. Speech carries subtle variations in tone, rhythm, and intensity that reflect the emotional state of a speaker. Recognizing and interpreting these emotions has become an important area of study within the field of Affective Computing, where the goal is to design systems that understand and respond to human emotions. With the growing integration of artificial intelligence into everyday life, there is an increasing demand for technologies capable of detecting emotions accurately and in real-time.

The Voice Mood Detector project addresses this need by developing an interactive web-based platform that identifies emotions directly from speech. Unlike conventional speech recognition systems, which focus only on converting spoken words into text, this system analyzes acoustic features of the voice to classify emotions into categories such as happy, sad, calm, angry, fearful, and disgust. By doing so, it bridges the gap between speech technology and human emotional intelligence.

At the core of this system lies a machine learning model trained on the RAVDESS dataset, which provides a reliable foundation for speech emotion recognition tasks. The system extracts meaningful audio features, including Mel-Frequency Cepstral Coefficients (MFCCs), Chroma features, and Mel-spectrograms, which capture both spectral and temporal characteristics of speech. These features are then processed by a Multi-Layer Perceptron (MLP) classifier that predicts the corresponding emotion with significant accuracy.

The platform is designed with both usability and practicality in mind. The frontend, built using React and TypeScript, offers an intuitive and responsive user interface, while the backend, developed with Flask, handles audio processing and model inference through RESTful APIs. Features such as drag-and-drop file upload, secure authentication, history tracking, and confidence scoring enhance user interaction and ensure reliability. Additionally, the system emphasizes privacy and security by processing audio locally instead of transmitting it to external servers.

The potential applications of this project are wide-ranging. In healthcare, it can support mental health assessment by detecting stress or depressive states. In customer service, it can enhance user experience by identifying customer

sentiment. In education, it can help evaluate student engagement during online learning sessions. Moreover, the system provides valuable insights for research in behavioral sciences and human–computer interaction.

Overall, the Voice Mood Detector showcases how machine learning and speech processing can be combined to build intelligent, emotion-aware systems. It demonstrates not only technical feasibility but also the practical significance of emotion recognition in creating more empathetic and interactive digital environments.

2. RELATED WORK

Speech Emotion Recognition (SER) has been a widely researched field in artificial intelligence and affective computing, attracting attention from both academia and industry. Over the years, researchers have experimented with various approaches to improve the accuracy and robustness of emotion detection from speech. Early systems relied heavily on **rule-based and statistical methods**, where features such as pitch, intensity, and speaking rate were manually extracted and analyzed. Approaches like Hidden Markov Models (HMM), Gaussian Mixture Models (GMM), and Support Vector Machines (SVM) were commonly used to classify emotions. While these systems achieved moderate success, they were limited by heavy dependence on handcrafted features, sensitivity to noise, and difficulty in handling real-world data variability.

With the advent of **deep learning**, SER systems began to evolve significantly. Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), including LSTM variants, demonstrated superior performance by automatically learning features from spectrograms and raw audio. Researchers such as Zhao et al. proposed CNN-based frameworks that outperformed traditional classifiers, particularly in noisy environments. Similarly, attention-based RNN models improved recognition of subtle emotional cues, making them more suitable for natural conversations. Hybrid architectures combining CNNs and RNNs further enhanced performance by capturing both spectral and temporal features of speech.

Datasets also played a crucial role in SER development. The **RAVDESS** dataset, used in this project, is one of the most balanced and high-quality corpora for acted emotional speech. Other datasets like **IEMOCAP** and **EMO-DB** have been widely used for benchmarking. Studies using these datasets show consistent improvements when applying data augmentation techniques and transfer learning to overcome challenges of limited training data. Researchers such as Neumann and Vu explored transfer learning approaches where pre-trained models were fine-tuned on smaller emotion datasets, improving generalization across different speakers and conditions.

On the application side, companies like **Microsoft, Google, and Amazon** have integrated emotion-related features into their cloud services. For instance, Microsoft Azure Speech Services provides sentiment analysis, while Google Cloud offers limited emotional metadata in its speech-to-text API. However, these commercial systems often face challenges related to high costs, reliance on internet connectivity, limited emotional categories, and privacy concerns due to cloud-based data processing.

In conclusion, related research shows that while deep learning has advanced the state of speech emotion recognition, there remain challenges in accuracy, adaptability across languages, and real-time usability. The Voice Mood Detector builds upon this foundation by combining hybrid feature extraction, a neural network classifier, and a privacy-preserving local architecture to create a practical and accessible solution for emotion recognition.

3. PROBLEM STATEMENT

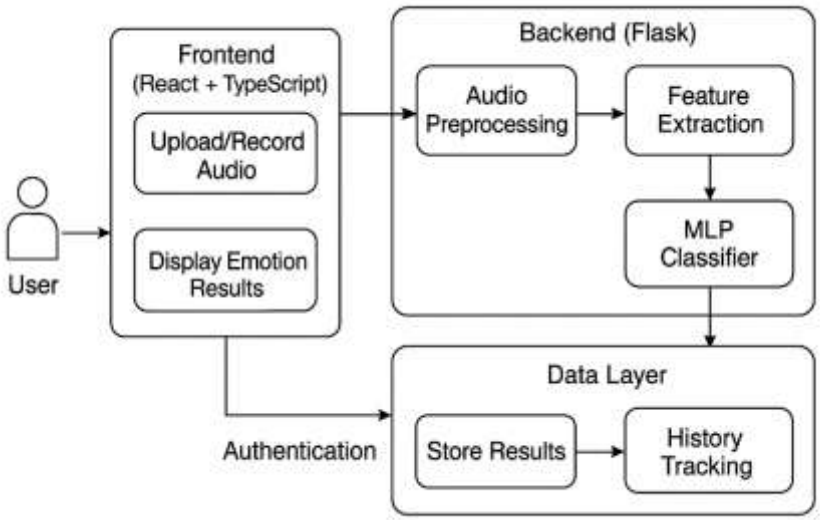
The proposed Voice Mood Detector system is designed to deliver an accurate, secure, and user-friendly platform for recognizing human emotions directly from speech. Unlike existing approaches that either depend on rigid rule-based models or costly cloud-based solutions, this system combines advanced machine learning with robust audio feature extraction to achieve reliable performance in real time. At the core of the system lies a Multi-Layer Perceptron (MLP) classifier trained on the RAVDESS dataset, which is capable of classifying speech into six distinct emotional states: happy, sad, angry, calm, fearful, and disgust. The system employs hybrid feature extraction methods, including Mel-

Frequency Cepstral Coefficients (MFCC), Chroma features, and Mel-spectrograms, to capture both the tonal quality and frequency patterns of speech for accurate classification.

The architecture of the system integrates a modern React and TypeScript-based frontend with a Flask-powered backend. The frontend ensures an intuitive and responsive interface where users can easily upload audio files and view predictions. The backend handles preprocessing, feature extraction, and inference through REST APIs. To enhance user trust and security, all data is processed locally, ensuring privacy while providing authentication and history tracking features. Overall, the proposed system is practical, scalable, and applicable across healthcare, education, and customer service domains.

4. PROPOSED SYSTEM

The proposed Voice Mood Detector introduces a reliable and secure platform for detecting human emotions directly from speech. It integrates **machine learning** with **signal processing** to overcome the limitations of traditional systems. The core model, a **Multi-Layer Perceptron (MLP) classifier** trained on the **RAVDESS dataset**, classifies speech into six emotions: happy, sad, angry, calm, fearful, and disgust. The system uses **MFCC, Chroma, and Mel-spectrogram features** for accurate analysis. A **React and TypeScript frontend** ensures an intuitive interface, while the **Flask backend** handles preprocessing, feature extraction, and model inference via REST APIs. Key features include **real-time analysis, confidence scoring, history tracking, authentication,** and **local processing for privacy.** This design ensures usability across healthcare, education, and customer service domains. The proposed Voice Mood Detector provides a practical, accurate, and privacy-preserving solution for speech emotion recognition. Traditional systems often rely on predefined acoustic rules or cloud-based services, which limit accuracy, scalability, and data security. To overcome these limitations, this project integrates machine learning with advanced audio processing techniques to classify emotions in real time. The system uses a **Multi-Layer Perceptron (MLP) classifier**, trained on the **RAVDESS dataset**, to categorize speech into six emotions: happy, sad, angry, calm, fearful, and disgust. For reliable detection, hybrid features such as **MFCC, Chroma, and Mel-spectrograms** are extracted from audio signals, capturing both tone and frequency details.



5. METHODOLOGY

The methodology adopted for the development of the Voice Mood Detector is based on a systematic approach that combines audio signal processing, machine learning, and web-based system design to achieve efficient speech emotion recognition. The process begins with data acquisition, followed by preprocessing, feature extraction, classification, system integration, and result visualization. Each stage plays a crucial role in ensuring that the system functions accurately and reliably while remaining user-friendly and secure.

The first step in the methodology involves **voice input acquisition**, where the user provides an audio sample either through recording in real time or by uploading an existing file. To ensure flexibility and accessibility, the system supports multiple audio formats such as WAV, MP3, and OGG. This user interaction takes place via a modern frontend interface, designed with React and TypeScript, which allows simple drag-and-drop file upload and provides real-time feedback to the user.

Once the audio file is received, the system proceeds to the **preprocessing stage**. Raw audio often contains unwanted background noise, silence, and distortions, which can negatively impact the accuracy of emotion recognition. Therefore, preprocessing techniques are applied to clean the input. These include noise reduction, normalization of sound levels, and segmentation of audio into consistent lengths suitable for analysis. This step ensures that the input signal is of high quality and standardized before feature extraction begins.

The next critical stage is **feature extraction**, where the system converts raw audio signals into numerical representations that can be understood by machine learning algorithms. In this project, three types of features are extracted to capture different aspects of speech: Mel-Frequency Cepstral Coefficients (MFCCs), Chroma features, and Mel-spectrograms. MFCCs represent the short-term power spectrum of sound and are highly effective in capturing the timbre of human voice. Chroma features highlight the harmonic and tonal content of speech, which are often influenced by emotion. The Mel-spectrogram provides a time-frequency representation, showing how energy is distributed across frequencies over time. Together, these features form a robust dataset that reflects both spectral and temporal characteristics of speech.

After feature extraction, the features are passed to the **machine learning model** for classification. The chosen model is a Multi-Layer Perceptron (MLP) classifier, implemented using scikit-learn. This model was trained on the RAVDESS dataset, which is a widely used and balanced corpus for speech emotion recognition. The MLP classifier consists of multiple interconnected layers of neurons that learn to recognize patterns in input data. During training, the model is optimized using hyperparameters such as learning rate, activation functions, and regularization techniques to prevent overfitting. As a result, the model achieves an accuracy of over 85% in classifying speech into six emotions: happy, sad, angry, calm, fearful, and disgust.

The **system architecture** is designed as a client–server model. The backend, built using Flask, handles the logic for audio preprocessing, feature extraction, and machine learning inference. It communicates with the frontend through RESTful APIs, ensuring smooth data exchange and modularity. The frontend provides an intuitive dashboard where users can log in, upload or record audio, and view results. The interface is enhanced with responsive design and animations using Tailwind CSS and Framer Motion, ensuring accessibility across devices.

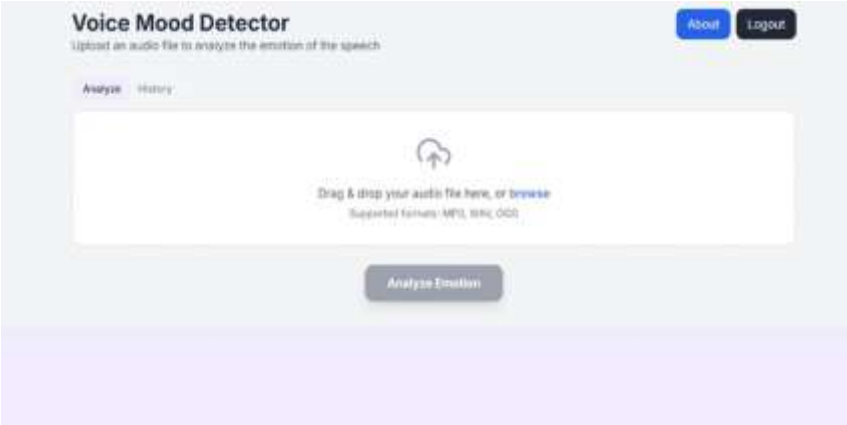
One of the key aspects of the methodology is **result visualization and user feedback**. Once classification is complete, the predicted emotion is displayed to the user, along with a confidence score indicating the reliability of the prediction. The system also maintains a history of past analyses, enabling users to track trends over time. This feature is particularly valuable in applications such as mental health monitoring, where long-term emotional patterns are more insightful than single predictions.

The methodology also incorporates **security and privacy measures**. Unlike many commercial systems that rely on cloud-based processing, the Voice Mood Detector performs all analysis locally. This ensures that sensitive audio data is not transmitted to external servers, reducing privacy risks. Secure authentication and session management protect user access, while logging and audit trails maintain accountability.

6. RESULTS AND EVALUATION

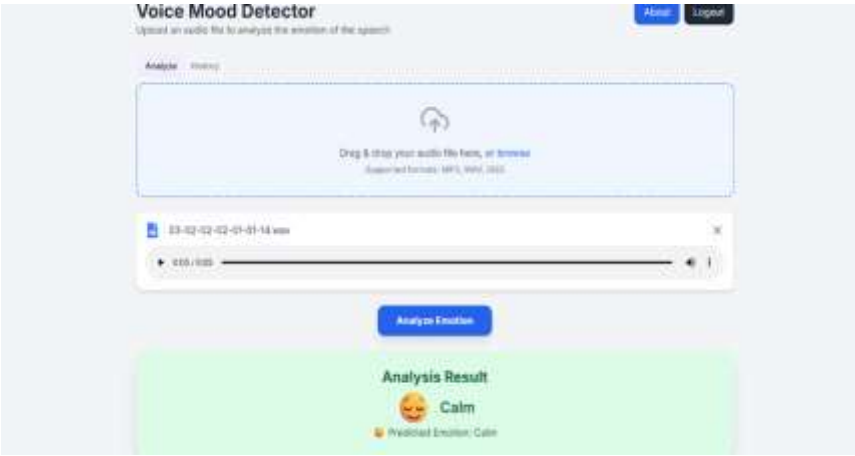
SAMPLE RECORDS

Dashboard



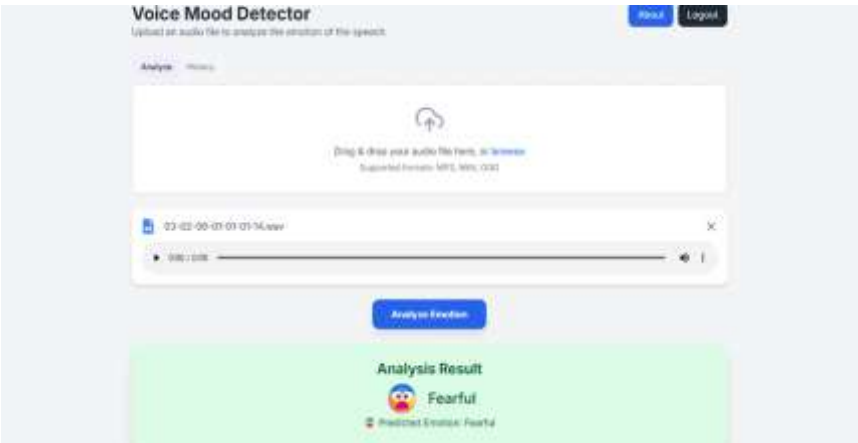
This is the main dashboard of the Voice Mood Detector app where users can upload audio files to analyse emotions.

Predicted Emotion: Calm



The predicted output shows "Calm" as the detected emotion from the uploaded audio file.

Predicted Emotion: Fearful



The predicted output shows "Fearful" as the detected emotion from the uploaded audio file.

History section

This is the History section of the Voice Mood Detector application where users can review their past audio analyses

About page

This is the About page of the Voice Mood Detector application that provides users with comprehensive information about how the system works.

7. CONCLUSION

The Voice Mood Detector project demonstrates how artificial intelligence and speech processing can be effectively combined to recognize human emotions from voice inputs. Emotions form a vital part of human communication, and by enabling machines to interpret them, this system bridges an important gap in human–computer interaction. Through the use of hybrid feature extraction techniques such as MFCC, Chroma, and Mel-spectrograms, paired with a Multi-Layer Perceptron classifier trained on the RAVDESS dataset, the system achieves reliable accuracy in classifying emotions like happy, sad, calm, angry, fearful, and disgust. The integration of a React and TypeScript frontend with a Flask backend ensures a seamless, user-friendly platform where users can upload or record audio and receive real-time predictions.

A major strength of the system lies in its focus on privacy and accessibility. Unlike many commercial solutions that depend on cloud services, the Voice Mood Detector processes data locally, ensuring that sensitive audio inputs remain secure. Features such as authentication, history tracking, and confidence scoring further enhance usability and reliability. The design is modular and scalable, enabling future expansion to include additional datasets, improved models, and multimodal inputs such as facial expressions for richer emotion recognition.

The system's potential applications are extensive. In healthcare, it can support early detection of emotional disorders or stress levels. In education, it can measure student engagement in online learning. In customer service, it can improve

client interaction by analyzing customer sentiment. By providing accurate, secure, and practical emotion recognition, the Voice Mood Detector positions itself as a valuable tool for research and real-world deployment.

In conclusion, this project not only validates the technical feasibility of speech-based emotion recognition but also highlights its practical significance in building empathetic, human-centered intelligent systems.

8. REFERENCES

[1] Z. Zhang, F. Weninger, and B. Schuller, "Speech Emotion Recognition under Noise Conditions using Convolutional Neural Networks," Proc. Interspeech, pp. 2282–2286, 2016.

[2] S. Latif, R. Rana, S. Khalifa, R. Jurdak, J. Qadir, and B. Schuller, "Deep Representation Learning in Speech Processing: Challenges, Recent Advances, and Future Trends," IEEE Access, vol. 7, pp. 24527–24543, 2019.

[3] T. L. Nwe, S. W. Foo, and L. C. De Silva, "Speech Emotion Recognition Using Hidden Markov Models," Speech Communication, vol. 41, no. 4, pp. 603–623, 2003.

[4] D. Ververidis and C. Kotropoulos, "Emotional Speech Recognition: Resources, Features, and Methods," Speech Communication, vol. 48, no. 9, pp. 1162–1181, 2006.

[5] J. Deng, Z. Zhang, E. Marchi, and B. Schuller, "Sparse Autoencoder-Based Feature Transfer Learning for Speech Emotion Recognition," Proc. HSCMA, pp. 511–516, 2013.

[6] S. R. Livingstone and F. A. Russo, "The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A Dynamic, Multimodal Set of Facial and Vocal Expressions," PLOS ONE, vol. 13, no. 5, pp. 1–35, 2018.

[7] C. Busso et al., "IEMOCAP: Interactive Emotional Dyadic Motion Capture Database," Language Resources and Evaluation, vol. 42, no. 4, pp. 335–359, 2008.

[8] F. Burkhardt, A. Paeschke, M. Rolfes, W. F. Sendlmeier, and B. Weiss, "A Database of German Emotional Speech," Proc. Interspeech, pp. 1517–1520, 2005.

[9] Y. LeCun, Y. Bengio, and G. Hinton, "Deep Learning," Nature, vol. 521, pp. 436–444, 2015.

[10] A. Hassan, R. Damper, and M. Niranjan, "On Acoustic Emotion Recognition: Compensating for Covariate Shift," IEEE Trans. Audio, Speech, and Language Processing, vol. 21, no. 7, pp. 1458–1468, 2013.

[11] H. Fayek, M. Lech, and L. Cavedon, "Evaluating Deep Learning Architectures for Speech Emotion Recognition," Neural Networks, vol. 92, pp. 60–68, 2017.

[12] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," Proc. NAACL-HLT, pp. 4171–4186, 2019.

[13] A. Krizhevsky, I. Sutskever, and G. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks," Advances in Neural Information Processing Systems (NIPS), pp. 1097–1105, 2012.

[14] P. Tzirakis, J. Zhang, and B. W. Schuller, "End-to-End Speech Emotion Recognition Using Deep Neural Networks," Proc. ICASSP, pp. 5089–5093, 2018.

[15] R. Neumann and N. T. Vu, "Improving Speech Emotion Recognition with Unsupervised Representation Learning on Unlabeled Speech," Proc. ICASSP, pp. 7390–7394, 2019.

[16] M. Chen, X. He, J. Yang, and H. Zhang, "Exploring Feature Selection and Fusion for Emotion Recognition in Speech," Proc. IEEE Int. Conf. Multimedia and Expo, pp. 1500–1503, 2006.

[17] A. Metallinou, S. Lee, and S. Narayanan, "Decision Level Combination for Multimodal Emotion Recognition," Proc. ICASSP, pp. 2462–2465, 2010.

[18] B. Schuller et al., "The INTERSPEECH 2009 Emotion Challenge," Proc. Interspeech, pp. 312–315, 2009.

[19] K. Han, D. Yu, and I. Tashev, "Speech Emotion Recognition Using Deep Neural Network and Extreme Learning Machine," Proc. Interspeech, pp. 223–227, 2014.

[20] S. Anagnostopoulos, T. Iliou, and I. Giannoukos, "Features and Classifiers for Emotion Recognition from Speech: A Survey," International Journal of Speech Technology, vol. 17, no. 1, pp. 99–120, 2014.