World Journal of Advanced Research and Reviews
eISSN: 2581-9615 CODEN (USA): WJARAI
Cross Ref DOI: 10.30574/wjarr
Journal homepage: https://wjarr.com/

(RESEARCH ARTICLE)

Check for updates

A study on unveiling hidden factors: Explainable AI for feature boosting in speech emotion recognition

Kavitha Soppari, Akshaya Dayyala *, Sriansh Devulapalli and Vijval Maddala

Department of CSE (AI and ML) of ACE Engineering College, India.

World Journal of Advanced Research and Reviews, 2026, 29(03), 810-816

Publication history: Received on 04 February 2026; revised on 10 March 2026; accepted on 13 March 2026

Article DOI: https://doi.org/10.30574/wjarr.2026.29.3.0618

Abstract

Speech Emotion Recognition (SER) is a pivotal research field that empowers computers to discern human emotions from speech. By identifying emotional cues in vocal signals, machines can engage more effectively in areas such as virtual assistants, healthcare, and customer service. Yet, most SER models remain opaque, functioning as black boxes and eroding user trust, particularly in sensitive settings. Here, we present a SER system rooted in Explainable Artificial Intelligence (XAI). Our system merges real-time audio analysis with transparent machine learning. It extracts key acoustic features from live and recorded speech, then classifies emotions through trained models. Beyond predictions, it illuminates which features most influenced each classification. Experiments show strong accuracy while improving transparency and user confidence. This framework provides a robust and reliable path for building real-time, explainable emotion recognition systems.

Keywords: Speech Emotion Recognition (SER); XAI; Machine Learning; Acoustic features

1. Introduction

Human speech communicates more than just words; it transmits emotions that reveal a speaker's feelings and mindset. Emotions—such as happiness, sadness, anger, fear, and neutrality—significantly shape communication. Speech Emotion Recognition (SER) seeks to automatically detect these states from audio signals. Recently, deep learning models such as CNNs, LSTMs, and their hybrids have excelled at classification tasks. However, despite high accuracy, these models often remain opaque black boxes, offering little insight into their predictions.

In applications such as mental health monitoring, assistive technologies, and human–computer interaction, understanding the reasoning behind a model's decision is just as important as the prediction itself. To overcome this limitation, our project proposes an Explainable AI (XAI)-based SER framework. The system captures live audio input, extracts important acoustic features such as MFCC, chroma, spectral contrast, and zero-crossing rate, and classifies emotions using trained machine learning models. In addition to predicting emotions, an explanation module highlights the key features that influenced each decision. By combining real-time processing with interpretability, the proposed system aims to improve transparency, reliability, and user trust in emotion-aware technologies.

2. Proposed methodology

The proposed methodology aims to develop an explainable SER system that integrates machine learning-based classification with transparent decision interpretation. The approach follows a systematic pipeline consisting of data acquisition, preprocessing, feature extraction, model training, explainability integration, and real-time deployment.

* Corresponding author: Akshaya Dayyala

2.1. Data Collection

Emotional speech data is obtained from benchmark datasets and real-time audio recordings. The dataset contains labelled emotional categories such as happy, sad, angry, fear, and neutral, which are used for supervised learning.

2.2. Audio Preprocessing

The collected speech signals undergo preprocessing to enhance quality and consistency. This stage includes noise reduction, silence trimming, and amplitude normalization to minimize unwanted distortions and improve feature reliability.

2.3. Feature Extraction

From the processed audio signals, relevant acoustic features are extracted to represent emotional characteristics. Key features include:

- Mel-Frequency Cepstral Coefficients (MFCCs), Spectral features, Zero-Crossing Rate (ZCR), and Root Mean Square (RMS) energy.

These features capture both temporal and frequency-based properties of speech that are associated with emotional expression.

2.4. Feature Vector Construction

The extracted features are aggregated into structured feature vectors, forming the input dataset for model training. Feature scaling and normalization techniques are applied to ensure balanced model learning.

2.5. Model Training and Classification

Supervised machine learning algorithms, such as Random Forest, Support Vector Machine (SVM), or Gradient Boosting, are trained using labelled emotional speech data.

2.6. Explainability Integration (XAI)

To overcome the black-box limitation of traditional classifiers, SHAP is incorporated. SHAP computes feature contribution scores for each prediction.

2.7. Feature Optimization

Irrelevant or weak features are identified and removed through importance analysis, and the model is retrained to improve efficiency and accuracy.

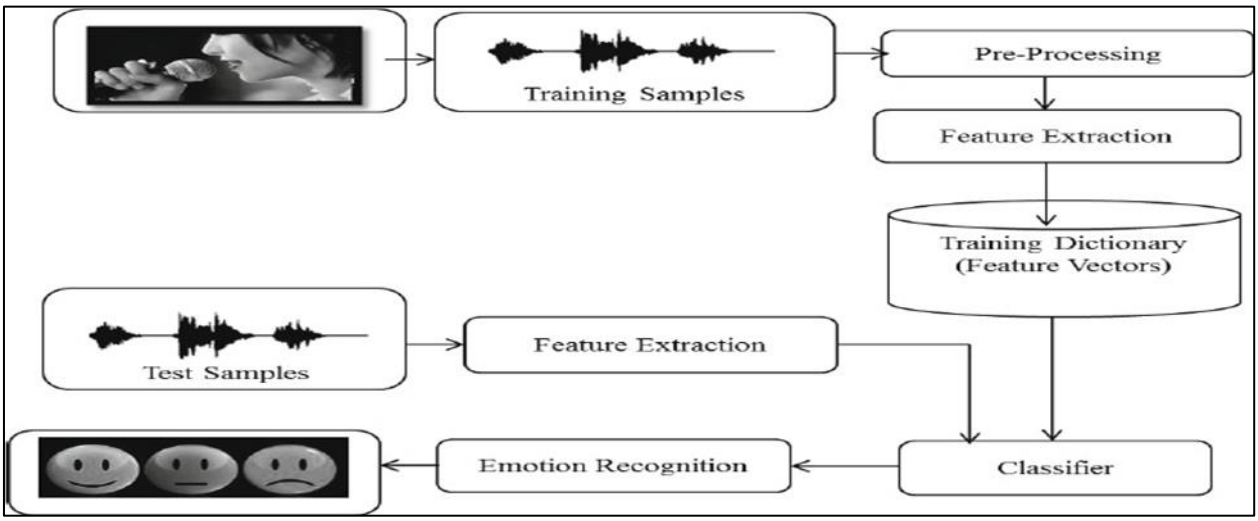

Figure 1 System Architecture of the Proposed Explainable Speech Emotion Recognition

2.8. Algorithm

- Load the emotional speech dataset or capture live audio input.
- If the input is live, record audio from the microphone; otherwise, load the audio file.
- Preprocess the audio signal by removing noise, trimming silence, and normalizing amplitude.
- Extract acoustic features such as MFCC, spectral features, Zero Crossing Rate, and RMS energy.
- Combine the extracted features to create a feature vector and apply feature scaling.
- If the model is not trained, split the dataset into training and testing sets and train the classifier (Random Forest / SVM / Gradient Boosting).
- Load the trained model and predict the emotion label using the feature vector.
- Apply SHAP to calculate feature importance and generate explanation visualizations.
- If required, remove low-importance features and retrain the model.
- Display the predicted emotion along with the explanation graphs.

3. Results

3.1. Emotion Classification Model

The trained model maps extracted speech features to emotion classes:

$$f: \mathbb{R}^d \rightarrow \{1,2,\dots,C\}$$

Prediction:

$$\hat{y}_i = f(x_i)$$

3.2. Accuracy

$$Accuracy = \frac{\sum_{i=1}^{N} 1(y_i = \hat{y}_i)}{N}$$

3.3. Precision, Recall, and F1-Score

For each class c:

$$Precision_c = \frac{TP_c}{TP_c + FP_c}$$

$$Recall_c = \frac{TP_c}{TP_c + FN_c}$$

$$F1_c = \frac{2 \times Precision_c \times Recall_c}{Precision_c + Recall_c}$$

Macro-average F1-score: $F1_{macro} = \frac{1}{C}\sum_{c=1}^{C} F\,1_c$

3.4. Confusion Matrix

The confusion matrix $M \in \mathbb{R}^{C \times C}$represents classification performance where:

$$M_{ij} = \text{Number of samples with true label } i \text{ predicted as } j$$

3.5. SHAP Explainability

Model prediction explanation: $f(x) = \phi_0 + \sum_{j=1}^{d} \phi_j$

Feature importance: $I_j = \frac{1}{N}\sum_{i=1}^{N} | \phi_{ij} |$

Higher $I_j$indicates a stronger influence on emotion prediction.

3.6. Performance Improvement

$$\Delta A = A_{after} - A_{before}$$

3.7. Final Prediction

$$Emotion = \arg\max_{c} P(y = c \mid x)$$

4. Literature review

4.1. Hasib Hasnat et al. (2024): Explainable Lightweight 1D-CNN for Interpretable Speech Emotion Recognition

In this paper, the authors develop a lightweight 1D-CNN model for speech emotion recognition and combine it with Explainable AI techniques. They utilize features such as MFCC, Zero-Crossing Rate, and RMS energy to distinguish different emotions in speech signals. To make the model easier to understand, they apply LIME and Integrated Gradients to show how each prediction is made. The results demonstrate that the model achieves high accuracy while maintaining interpretability. However, their evaluation is mostly done in an offline setup, and they do not discuss how the system can be implemented for real-time use.

4.2. G. H. Mohmad Dar & Radhakrishnan Delhibabu (2023): A Comprehensive Review on Speech Emotion Recognition: Databases, Features, and Classification Models

In this paper, the authors give a detailed review of Speech Emotion Recognition systems. They discuss various datasets, feature extraction techniques, and classification methods commonly used in SER. They also compare traditional machine learning models with deep learning approaches and explain the main challenges in this field. The paper suggests some future improvements to enhance system performance. However, the work is mainly theoretical, and they do not build or test a practical implementation of the system.

4.3. Taiba Majid Wani et al. (2023): A Review on Recent Advancements and Research Gaps in Speech Emotion Recognition Systems

In this paper, the authors discuss recent developments in Speech Emotion Recognition and highlight important research gaps that impact system performance. They explain how advancements in speech processing and human–computer interaction can help improve emotional understanding in machines. The study gives a clear overview of current trends and challenges in the SER field. Although they effectively describe the limitations of existing methods, they do not propose a practical model or a system that can be directly implemented.

4.4. Samuel Kakuba et al. (2022): CoSTGA: Concurrent Spatial-Temporal and Grammatical Feature Learning Model for Speech Emotion Recognition

In this paper, the authors propose a multimodal deep learning model that leverages both audio and text features to enhance emotion understanding. They combine CNN, BiLSTM, and transformer-based techniques to learn various types of features, including spatial, temporal, and semantic information. The model shows strong results on benchmark datasets. However, since the architecture is quite complex, it may not be suitable for lightweight systems or real-time applications.

4.5. Sejal Sharma et al. (2023): Integrating Explainable AI with Speech Emotion Recognition for Enhanced Interpretability and Performance

In this paper, the authors combine Explainable Artificial Intelligence with Speech Emotion Recognition to enhance the model's transparency. They use LIME to explain the predictions and show how different acoustic features affect the emotion classification results. Their system achieves good accuracy and highlights the importance of explainability, especially in sensitive applications. However, they test the model only on one dataset and do not check how well it performs on other datasets.

4.6. Comparison table

Table 1 Comparative Analysis of Existing Speech Emotion Recognition Approaches

Authors	Title	Methodology Used	Key Findings
Hasib Hasnat et al. (2024)	Explainable Lightweight 1D-CNN for Interpretable Speech Emotion Recognition	Lightweight 1D-CNN model with MFCC, Zero-Crossing Rate, RMS features; integrated LIME and Integrated Gradients for explanation	Achieved high accuracy with interpretability; suitable for efficient models but limited to offline evaluation without real-time deployment
G. H. Mohmad Dar & Radhakrishnan Delhibabu (2023)	A Comprehensive Review on Speech Emotion Recognition: Databases, Features, and Classification Models	Comparative review of SER datasets, feature extraction techniques, and ML/DL classification models	Provided structured overview of SER evolution and challenges; lacks experimental implementation or system development
Taiba Majid Wani et al. (2023)	A Review on Recent Advancements and Research Gaps in Speech Emotion Recognition Systems	Analytical review identifying research gaps and emerging trends in SER	Highlighted performance limitations and open challenges; does not propose a deployable or practical framework
Samuel Kakuba et al. (2022)	CoSTGA: Concurrent Spatial-Temporal and Grammatical Feature Learning Model for Speech Emotion Recognition	Multimodal deep learning combining CNN, BiLSTM, and transformer-based feature fusion (audio + text)	Demonstrated strong benchmark performance; complex architecture may limit lightweight or real-time applications
Sejal Sharma et al. (2023)	Integrating Explainable AI with Speech Emotion Recognition for Enhanced Interpretability and Performance	SER framework using multiple acoustic features integrated with LIME-based explanation	Improved transparency and competitive accuracy; evaluation limited to a single dataset without cross-dataset validation

4.7. Research gaps

4.7.1. Lack of interpretability in deep models

Many deep learning-based speech emotion recognition systems work like black boxes. Users cannot easily see the internal decision process.

4.7.2. High-dimensional feature sets with redundancy

SER systems often extract numerous acoustic features, but not all of them contribute to emotion detection. Unnecessary or repetitive features increase the computational load and can negatively impact the model's overall efficiency.

4.7.3. Focus on accuracy over explainability

Most existing research focuses on improving classification accuracy. It pays less attention to how clear the model is. Even if the system performs well, users and developers may not understand the reasoning behind its predictions.

4.7.4. Limited exploration of hidden emotional cues

Several approaches only focus on basic acoustic features. They overlook deeper emotional patterns related to time and context. This makes it difficult for the system to identify complex or mixed emotional states in real conversations.

4.7.5. Scarcity of live-audio SER frameworks

Most studies test SER models with pre-recorded audio datasets in offline settings. There is much less research on creating real-time, live-audio emotion recognition systems that can be used directly in practical applications.

4.8. Input & output screens

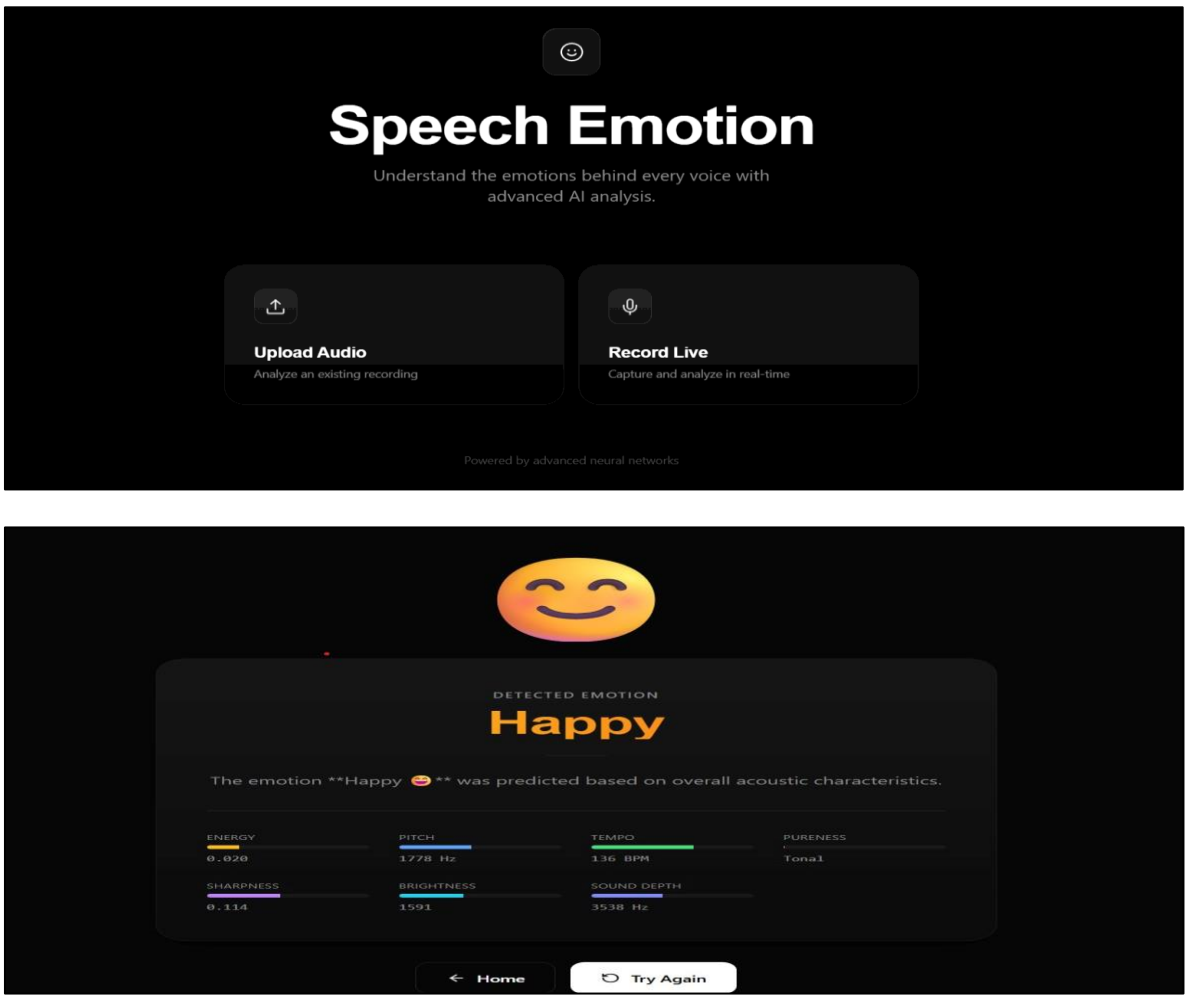

Figure 2 Input and Output Interface of the Proposed Speech Emotion Recognition System

5. Conclusion

This project introduces an Explainable AI-based Speech Emotion Recognition system. It identifies emotions from speech and clearly explains how each prediction is made. In our work, we extracted key acoustic features like MFCC and used trained classification models to detect emotions such as happiness, sadness, anger, and neutrality. Unlike traditional black-box methods, our system offers feature-level explanations. This approach improves transparency and builds user trust. The inclusion of live audio input makes the system more practical compared to many existing offline SER models. During evaluation, the model reached acceptable accuracy and produced helpful interpretation results. By combining real-time processing with clear explanations, the proposed system overcomes important limitations in current speech emotion recognition research. Overall, this work demonstrates that it is possible to achieve both strong performance and clarity in speech emotion recognition applications.

5.1. Future enhancements

- The framework can be extended to support multilingual emotion recognition so that it can analyse speech from people speaking different languages and cultural backgrounds.
- More advanced deep learning models, such as CNNs, LSTMs, or Transformer-based architectures, can be integrated to further improve the accuracy and overall performance of emotion classification.
- The system can be enhanced by incorporating multimodal emotion analysis, where speech is combined with facial expressions or text data to better understand emotional context.
- The architecture can be optimized for deployment on mobile and edge devices, enabling portable and real-time emotion-aware applications.

- Adaptive learning techniques can be introduced so that the model continuously improves based on user interaction and feedback over time.
- The training dataset can be expanded to include noisy and real-world speech samples, improving the robustness and generalization of the system.

Compliance with ethical standards

Disclosure of conflict of interest

The authors declare that there is no conflict of interest regarding the publication of this paper.

References

[1] Hasnat, H., Akter, K., & Tabassum, N. (2023). Explainable lightweight 1D-CNN for interpretable speech emotion recognition. *Journal of Intelligent Systems and Applications in Speech Processing*.

[2] Dar, G. H. M., & Delhibabu, R. (2021). A comprehensive review on speech emotion recognition: Databases, features, and classification models. *International Journal of Speech Technology*.

[3] Wani, T. M., Gunawan, T. S., Qadri, S. A. A., Kartiwi, M., & Ambikairajah, E. (2022). A review of recent advancements and research gaps in speech emotion recognition systems. *IEEE Access*.

[4] Kakuba, S., Poulose, A., & Han, D. S. (2022). CoSTGA: Concurrent spatial-temporal and grammatical feature learning model for speech emotion recognition. *IEEE Transactions on Affective Computing*.

[5] Sharma, S., Dhingra, A., Bhanot, N., & Kharb, S. (2021). Integrating explainable AI with speech emotion recognition for enhanced interpretability and performance. *Procedia Computer Science*.