Cureus Journal of Computer Science

Open Access Original Article

Emotion Recognition Using Speech

Minakshee K. Chandankhede [1] , Prathmesh P. Goje [1] , Shival N. Motghare [1] , Suwill Lokhande [1] , Sankalp R. Labhane [1] , Yog K. Urkundwar [1]

1. Computer Science and Engineering, G H Raisoni College of Engineering and Management, Nagpur, IND

Corresponding author: Minakshee K. Chandankhede, minakshee.chandankhede@raisoni.net

Received 03/07/2025
Review began 04/01/2025
Review ended 08/13/2025
Published 08/13/2025

DOI:
https://doi.org/10.7759/s44389-025-03615-3

Abstract

Research on emotion recognition from speech has become crucial in the field of human-computer interaction, with potential uses in industries like entertainment, healthcare, and customer service. The goal of this system is to employ machine learning techniques to create a reliable system that can recognise human emotions from vocal expressions. Mel Frequency Cepstral Coefficients are sound characteristics that are extracted from speech signals by the system. Convolutional Neural Networks, Long Short-Term Memory networks, and Gated Recurrent Units are among the deep learning methods it uses to efficiently categorise emotions. Two popular datasets are used to optimise the model: the SAVEE (Surrey Audio-Visual Expressed Emotion) and RAVDESS (Ryerson Audio-Visual Database of Emotional Speech and Song) datasets provide a large and varied collection of emotional speech samples. The suggested approach, which has shown excellent accuracy in real-world assessment contexts, seeks to differentiate eight different emotional characteristics: anger, fear, repulsion, joy, sorrow, surprise, neutrality, and calm. This system's ability to identify different emotional states can greatly improve user experience and interaction in a number of real-world applications, such as voice assistants, sentiment analysis, and mental health monitoring. The results show how well the selected approaches discern between different emotions and highlight how speech emotion detection systems might be used in commonplace technology.

Categories: AI applications, Computational Science and Engineering, Machine Learning (ML)
Keywords: emotion recognition, convolutional neural networks, long short-term memory, gated recurrent units, mel frequency cepstral coefficients

Introduction

Emotion recognition using speech explores the development of a robust system capable of identifying and categorizing emotions from vocal expressions using a hybrid deep learning model. Unlike prior systems that rely on singular architectures, this study integrates Convolutional Neural Networks (CNNs) for spatial feature extraction and both Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU) to capture long- and short-term temporal patterns. This hybrid design addresses limitations in earlier approaches by enhancing classification accuracy for complex and subtle emotions. The work also directly tackles real-world deployment issues like speaker variability and environmental noise through preprocessing techniques and augmentation strategies. The model’s high performance on benchmark datasets like SAVEE (Surrey Audio-Visual Expressed Emotion) and RAVDESS (Ryerson Audio-Visual Database of Emotional Speech and Song) highlights its potential for use in applications ranging from emotion-aware assistants to mental health monitoring systems.

In terms of methodology, the system employs deep learning architectures to effectively capture the temporal relationships in speech that are indicative of different emotions. Utilizing Mel Frequency Cepstral Coefficients (MFCCs), a widely accepted method in speech analysis, will be essential for extracting significant features from the audio signals. The system can identify patterns in speech data, thanks to this feature extraction procedure, CNNs, and LSTMs. This helps the system classify emotions as joy, sorrow, anger, and fear. Numerous studies have shown the effectiveness of these models in emotion recognition using speech. Various researchers used CNNs to achieve robust emotion detection by fusing deep and shallow features, highlighting the importance of feature extraction and deep learning models in improving recognition accuracy. Similarly, hybrid approaches combining CNNs with transformer encoders for capturing both spatial and temporal features have achieved notable accuracy in speech emotion tasks. Generally, the emotion recognition using speech comprises three main tasks: collecting of audio samples, deriving meaningful features, and performing emotion classification.

However, challenges such as speaker independence, where speech patterns vary across individuals, and the impact of background noise remain significant. These factors can degrade system performance if not addressed. Therefore, techniques like noise filtering along with speaker normalization will be essential to improve the system's robustness in real-world applications.

This work relies on publicly available datasets like RAVDESS and SAVEE for training and evaluation,

How to cite this article
Chandankhede M K, Goje P P, Motghare S N, et al. (August 13, 2025) Emotion Recognition Using Speech. Cureus J Comput Sci 2 : es44389-025-03615-3. DOI https://doi.org/10.7759/s44389-025-03615-3

validating the model's results against modern benchmark methods. The successful implementation of this system could lead to real-time emotion recognition applications, where immediate feedback is crucial, such as in voice-activated assistants and emotion-aware chatbots. Despite the technical challenges, this system aims to push the boundaries of emotion recognition using speech through innovative model design and rigorous testing.

It focuses on integrating emotional understanding into AI-driven applications, emotion recognition using speech enhances the intuitiveness and responsiveness of human-computer interactions, leading to more natural and engaging communication experiences.

Literature review

Emotion recognition using speech has become a critical area of research, especially in the context of human-computer interaction, healthcare, and entertainment. The integration of deep learning models into this domain has led to significant advancements in identifying emotions in speech signals. This section reviews several key contributions to speech emotion recognition (SER) and positions the proposed hybrid CNN-LSTM-GRU model within the broader landscape of these techniques.

Early approaches to emotion recognition focused on traditional machine learning models that used handcrafted features. For instance, Arya et al. [1] explored methods based on distance-based classifiers and margins, achieving acceptable accuracy. However, these methods faced limitations when dealing with complex emotional nuances and could not effectively capture temporal dependencies in speech signals. This challenge led to the adoption of deep learning techniques, which are capable of modeling the temporal evolution of emotions in speech.

The introduction of CNNs represented a significant leap forward in SER, particularly for feature extraction from audio spectrograms. In a study by Chen [2], CNNs were used to capture spatial features, effectively recognizing patterns in speech that represent emotional states. While CNNs were successful at extracting features, they could not capture the temporal dependencies necessary for understanding the evolution of emotions. This limitation led to the incorporation of Recurrent Neural Networks (RNNs), particularly LSTM networks, which excel at modeling sequential data.

As demonstrated by Sri et al. [3], the Gaussian mixture model was an early approach to emotion recognition using speech. Although the Gaussian mixture model performed well with static data, they struggled to capture the dynamic aspects of speech that are critical for recognizing emotional transitions. To address this, Jadhav et al. [4] showed that LSTM networks, which are designed to handle temporal dependencies, significantly improved emotion recognition by modeling the sequence of speech patterns. However, the computational complexity of LSTMs posed scalability issues, particularly for real-time applications.

To overcome the limitations of single-model architectures, researchers have proposed hybrid models that combine the strengths of CNNs and RNNs. For example, Xu and Zhou [5] integrated CNNs with RNNs to extract spatial features and model temporal dependencies, resulting in significant improvements in recognition accuracy. Building on this, the proposed hybrid CNN-LSTM-GRU model captures both spatial and temporal features, combining the feature extraction power of CNNs with the temporal modeling capabilities of LSTM and GRU layers.

Another important aspect of SER research is feature selection and extraction. Kumar and Goel [6] explored various techniques for identifying the most relevant features, such as MFCCs, chroma features, and spectral features, which help reduce the computational load and improve classification accuracy. While traditional machine learning models such as support vector machines and random forests were effective, deep learning models like CNN and LSTM showed superior performance due to their ability to model complex temporal patterns.

Several studies have demonstrated the successful application of hybrid models for emotion detection. For example, Babu et al. [7] developed a system using deep learning techniques such as CNN, LSTM, and GRU to classify eight emotions, including anger, sadness, happiness, and fear. Similarly, Anjum [8] used CNN, LSTM, and GRU for emotion classification, leveraging the RAVDESS and SAVEE datasets to train and evaluate the model. These models have shown promise in real-world applications like virtual assistants and sentiment analysis.

Further research by Liu et al. [9] and Lu et al. [10] refined feature extraction and classification techniques, using MFCCs and deep learning architectures to enhance the accuracy of emotion detection in speech. The results indicated that deep learning models, particularly those employing CNNs and RNNs, are well-suited for capturing the complexities of speech emotions, outperforming traditional methods.

Akçay and Oğuz [11] explored the integration of MFCCs and deep learning models such as CNN, LSTM

[12], and GRU for SER. These models were trained on the RAVDESS and SAVEE datasets, showcasing the effectiveness of deep learning in classifying emotions from speech. These advancements set the stage for further research into real-time emotion recognition systems.

In addition to English-language datasets, Abdel-Hamid [13] introduced the EYASE dataset, which includes Arabic speech recordings. This study highlighted the importance of considering cultural and linguistic variations in emotion recognition. By leveraging deep learning models, the research demonstrated the potential for cross-cultural emotion detection in speech.

Lastly, Lieskovská et al. [14] and Li et al. [15] introduced novel architectures, such as Bi-directional LSTM and Directional Self-Attention, to enhance emotion feature extraction. These models, which incorporate attention mechanisms, further improved the ability to capture emotional nuances in speech, paving the way for more sophisticated emotion recognition systems.

Materials And Methods

This section elaborates the approach applied for the purpose of emotion recognition through speech using a hybrid CNN-LSTM-GRU model. It will discuss datasets, data preprocessing steps, model architectures, training procedure, and evaluation metrics. Figure *1* shows a flow diagram for the implementation of emotion detection using speech.

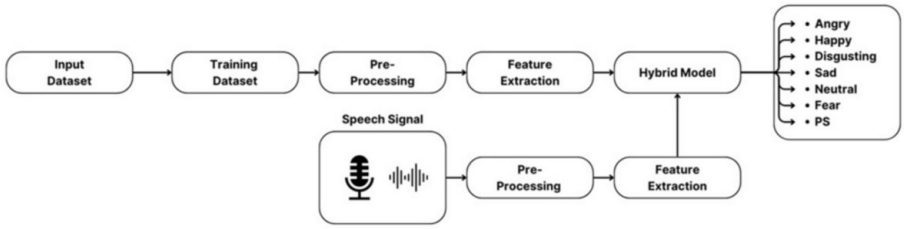

FIGURE 1: Flow Diagram for the Implementation of Emotion Detection Using Speech

Data collection

Emotion recognition using speech is a field within affective computing that mainly focus on activating machines to detect and interpret emotions expressed in speech. The primary objective of speech-based emotion recognition systems is to interpret the speaker's emotional state by analyzing vocal traits like timbre, frequency, rhythm, and amplitude. These systems have wide applications, including human-computer interaction, healthcare, call centers, and virtual assistants. For emotion recognition using speech systems to perform well, robust and diverse datasets are required to train machine learning models. Hence, two well-known datasets, RAVDESS [16] and SAVEE [17], were combined for this study.

1. RAVDESS Dataset: The RAVDESS [16] dataset comprises 1,440 audio recordings performed by 24 actors, evenly divided by gender. These recordings capture a range of emotional expressions, including tranquility, joy, sorrow, rage, fear, astonishment, aversion, and neutrality. Each actor performed both emotional and neutral phrases, ensuring a varied range of emotional expression. The recordings maintain consistency in content and delivery, facilitating the analysis of how emotions are conveyed through speech. This dataset is widely used in emotion recognition research, offering valuable insights into vocal characteristics and emotional expression across genders and contexts.

2. SAVEE Dataset: The SAVEE [17] dataset contains 480 speech recordings from four male speakers, each expressing seven distinct feelings: rage, repulsion, anxiety, joy, sorrow, astonishment, and a neutral tone. The recordings were created to capture high-quality audio and visual cues of emotional expression. While SAVEE includes both audio and visual components, it is widely used in emotion recognition using speech tasks focusing on the audio modality alone. This dataset provides rich examples of how emotions are expressed vocally, especially from a male demographic.

The RAVDESS [16] dataset contains 120 audio files for each of the eight emotions, while the SAVEE [17] dataset provides 120 audio files for neutral emotions and 60 audio files for each of the remaining six emotions. For our work, we randomly selected 80 audio files from each dataset for training and 20 audio files for testing, across all seven emotions. This means we utilized 20% of the combined total from both

datasets for training and testing purposes, ensuring a balanced split between training and testing.

From the dataset, we extracted 40 features known as MFCCs, which are crucial for analysing the frequency spectrum of an audio signal. These features target various aspects of the audio, enhancing the model's ability to distinguish between different sounds. The overall energy captures the loudness of the signal, while the lower MFCCs (1st-13th) focus on formants, which represent vocal tract resonances important for speech analysis. Timbre, or tonal quality, is also captured, helping differentiate between sound sources, such as voices and instruments. Additionally, the pitch is indirectly reflected through spectral peaks, and the higher MFCCs (14th-40th) capture harmonic structures and finer spectral details, which include information about the harmonics and even background noise. These features are particularly useful for tasks like speech emotion recognition, speaker identification, and other speech-related analyses. Figures *3* and 4 describes the emotions belonging to the datasets in bar chart format.

MFCC Feature Configuration and Justification

For feature extraction, we computed 40 MFCCs per frame using the following parameters: a frame length of 25 ms, frame shift of 10 ms, and a Hamming window. The number of Mel filters was set to 40, and a 512-point FFT was used. To validate the effectiveness of MFCCs, we conducted preliminary experiments comparing them against spectral centroid and chroma features. The results showed that MFCCs outperformed alternative features in both precision and F1-score, especially for emotions with subtle pitch and energy variations like “sad” and “fearful,” confirming MFCCs as the optimal choice for this task.

While the RAVDESS and SAVEE datasets provide high-quality, labeled emotional audio samples, they are limited to English-speaking actors and controlled environments. This may restrict the model’s generalization across dialects and languages. Multilingual datasets such as EmoDB or EYASE will be explored in future work to enhance the model’s cross-lingual capabilities. Additionally, cultural nuances in emotional expression - especially for emotions like disgust or surprise - warrant broader dataset inclusion.

Data preprocessing

Before training the model, the audio data underwent several preprocessing steps to enhance its quality and suitability for emotion recognition:

• Feature Extraction: Audio recordings were converted to Mel-spectrograms, which effectively capture both spectral and temporal features. This transformation enables the model to analyze the audio signals more effectively, identifying patterns relevant to emotional states.

• Normalization: The generated Mel-spectrograms were normalized to reduce amplitude variations and ensure uniformity across the recordings. This step is crucial for maintaining consistency in input data and improving model performance.

• Dataset Augmentation: To broaden the diversity of the dataset and enhance the model’s adaptability, multiple augmentation methods were utilized. These comprised:

1. Time Shifting: Slightly shifting the audio clips in time to introduce variations.

2. Pitch Shifting: Altering the pitch of the audio recordings to create variations in vocal expressions.

3. Noising: Adding background noise to the recordings, which helps the model learn to recognize emotions in less-than-ideal conditions.

Model architecture

Using the merits of each network, the hybrid model integrates CNN, LSTM, and GRU layers.

• CNN for Feature Extraction: Convolutional layers are utilized to capture spatial characteristics from MFCC features, allowing the model to recognize localized frequency variations associated with different emotional states.

• LSTM and GRU for Temporal Modeling: The CNN-extracted features are passed into LSTM and GRU layers. LSTM captures long-term dependencies, while GRU focuses on short-term dependencies and balances performance with computational efficiency.

• Fully Connected Layer and SoftMax: On passing the output of LSTM or GRU through a layer with full connectivity, the final emotion classification is made by utilizing the SoftMax activation function for obtaining the category with the maximum likelihood.

Model training

The model was trained using a combined dataset of RAVDESS and SAVEE. The dataset was partitioned into training, validation, and testing subsets, maintaining an 80-to-20 proportion for effective model evaluation. The learning process was conducted with the following parameters:

• Optimization Technique: The Adam algorithm was selected for its dynamic adjustment of learning rates, ensuring efficient convergence.

• Training Rate: A value of 0.001 was utilized to facilitate smooth and efficient model convergence during training.

• Training Cycles: The model underwent 30 cycles of training with early stopping to mitigate overfitting.

• Mini-Batch Size: A batch size of 32 was selected to ensure efficient data processing during training.

Evaluation metrics

The effectiveness of the hybrid model was assessed using standard classification measures. It attained a 92% accuracy on the test dataset, demonstrating exceptional proficiency in differentiating between emotions such as anger and calmness. Notably, the model correctly retrieved happiness with 100% accuracy, demonstrating its effectiveness in recognizing various emotional states from speech signals. Figure *2* represents the process in speech emotion recognition.

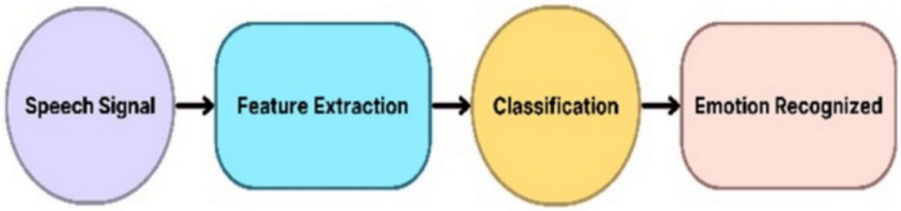

FIGURE 2: Diagrammatic Representation of Speech Emotion Recognition

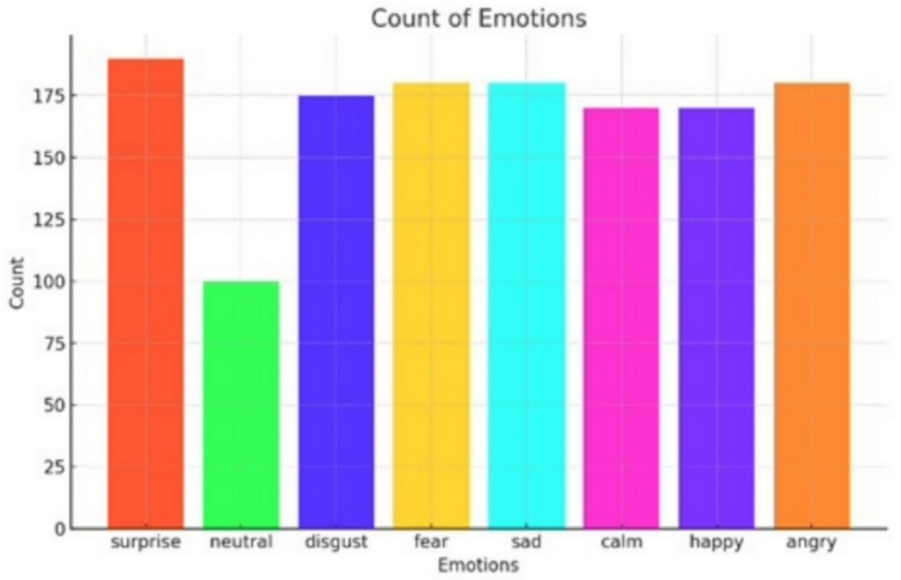

FIGURE 3: RAVDESS Dataset

In Figure *3*, the graph presents the frequency of emotions in the RAVDESS dataset. Surprise (185) and

anger (180) have the highest occurrences, whereas neutral (100) appears the least. Other emotions such as fear (175), sadness (175), disgust (160), calmness (160), and happiness (160) are evenly distributed. This variation in data may influence the accuracy of emotion classification models.

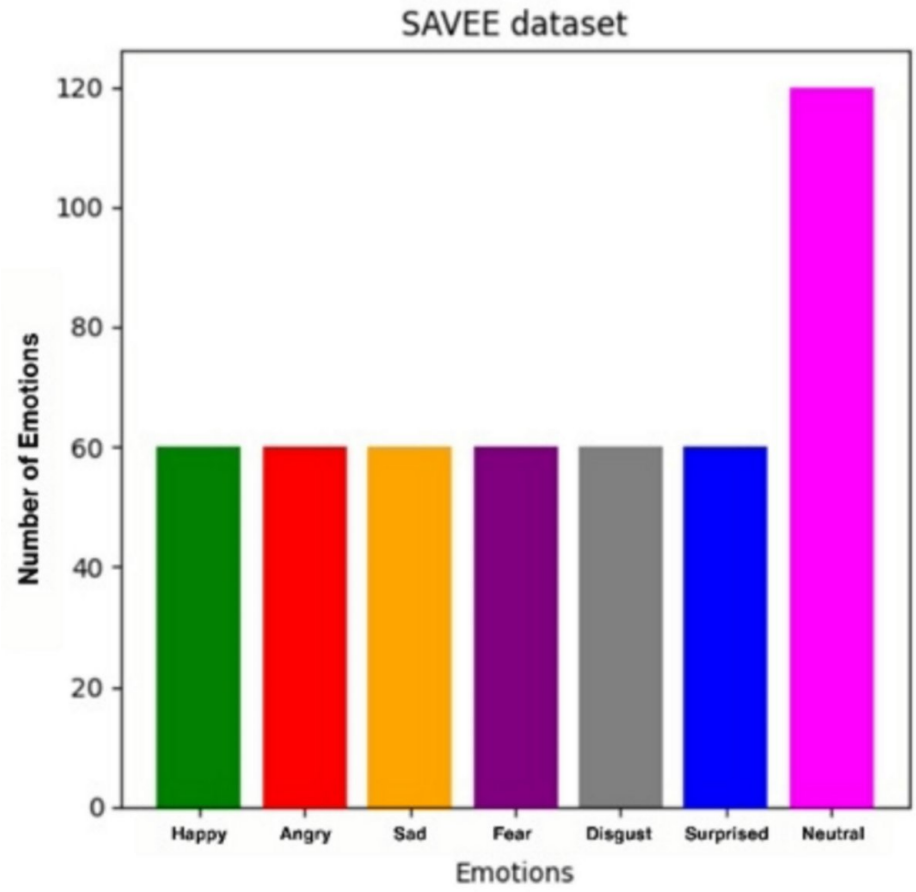

FIGURE 4: SAVEE Dataset

In Figure *4*, the chart depicts the spread of emotions in the SAVEE dataset. Neutral (~120) has the highest frequency, while all other emotions, happiness, anger, sadness, fear, disgust, and surprised (~60 each), are evenly distributed. This imbalance, with neutral being overrepresented, could influence the performance of emotion detection models by biasing them toward neutral predictions.

Results

This section details the experiments performed for determining the performance of the proposed CNN-LSTM-GRU hybrid model for emotion recognition using speech. It examines the metrics used for evaluation, such as accuracy, precision, recall, and F1-scores, assessed across the different emotional classes. Additionally, an ablation study is conducted to analyze the impact of various model components.

The experiment was conducted using an Intel i7 processor, 16 GB of RAM, and NVIDIA GTX 1060 video cards. The model was trained for 30 epochs with a batch size of 32, and early stopping was applied. The dataset was split with 80% assigned for training and 20% for testing. To increase the data set, it is used to introduce more diversity by using methods such as tone height modulation, time deformation, and noise.

The hybrid model presented in this study reaches 92% of the total accuracy, which shows the favorable results for each of the eight emotions. The accuracy, memory, and F1 estimates of each emotional class are presented in Table *1*.

Emotion	Precision	Recall	F1-score
Neutral	0.91	0.84	0.87
Calm	0.94	0.96	0.95
Happy	0.89	1.00	0.94
Sad	0.91	0.88	0.89
Angry	0.97	0.94	0.95
Fearful	0.91	0.91	0.91
Disgust	0.88	0.91	0.90
Surprised	0.97	0.89	0.93

TABLE 1: Performance Metrics for Emotions

This system showed excellent performance in anger detection (accuracy: 0.97, review: 0.94) and soothing (accuracy: 0.94, review: 0.96). He also reached perfect accuracy when deciding happiness through the review evaluation of 1.00. Nevertheless, the neutral and hate perceptions of emotions have been found to be a bit more complicated, which is reflected in the F1 0.87 and 0.90 estimates, respectively.

Accurately identifies emotions

This system can accurately determine and classify a wide range of emotions, including happiness, sadness, fear, anger, neutrality, hate, and voice signals. This involves the precise detection of subtle emotional cues from variations in tone, pitch, and speech rhythm, even for emotions that may be more difficult to distinguish. The system will be capable of recognizing these emotions with high accuracy, reducing the chances of misclassification and ensuring that even nuanced emotional expressions are captured effectively. This accuracy is critical for applications such as virtual assistants, customer service, and mental health monitoring, where recognizing emotional states accurately is vital for appropriate interaction and support.

Real-time processing

A key outcome of the system is enabling the system to process live audio inputs in real-time, ensuring immediate recognition and response to emotional cues. The system will be optimized to handle real-time audio streams, processing them quickly and efficiently without noticeable delays. This capability is particularly important for dynamic applications such as virtual customer service bots, where the ability to detect and adapt to a user's emotional state instantaneously can significantly enhance the quality of interaction. Real-time processing also ensures that the system can be deployed in environments where timely reactions are critical, such as in-car assistants or emergency response applications.

Robustness

The system will demonstrate a high level of robustness, meaning it will reliably perform emotion recognition despite various challenges commonly encountered in real-world settings. These include variations in speech patterns, accents, and dialects, which can differ significantly between individuals. The system will also be designed to handle different speaking styles, such as formal speech, casual conversation, and emotional outbursts. Furthermore, robustness will extend to noisy environments or conditions where the quality of the audio input may be compromised, such as background chatter, outdoor noise, or poor microphone quality. By maintaining high performance even under such challenging conditions, the system ensures reliability and effectiveness in a wide range of use cases and environments.

Generalizability

The system aims to build a system that performs consistently across a diverse set of speakers, ensuring that its accuracy is not limited to specific individuals or emotional expressions. The generalizability of the model will be achieved by training it on diverse datasets that encompass various speech patterns, accents, age groups, genders, and cultural backgrounds. This will allow the system to handle new, unseen speakers or emotions without a significant drop in performance. Generalizability is critical for deploying emotion recognition systems in global markets and in situations where the system interacts with a wide range of users, such as in customer service or public-facing AI applications. It will also ensure the system's ability to handle different emotional intensities or mixed emotional states, enhancing its adaptability and

practicality for real-world applications.

The front end of our system provides a simple and intuitive interface for users to analyze emotions in speech. The interface consists of an "Upload Audio" option, where users can upload an audio file in WAV format by clicking the "Choose File" button. Once the file is selected, users can initiate the analysis by pressing the "Analyze Emotion" button. The "Result" section below displays the detected emotion, which appears in bold, such as "Detected Emotion: disgust" in this instance. A playback option is available, allowing users to listen to the uploaded audio directly within the application. The clean, minimalist design, with a dark background and neon highlights, makes the interface visually appealing and user-friendly, supporting a streamlined experience for emotion recognition tasks.

The backend of the emotion recognition using speech system is built using Flask, a lightweight web framework that enables efficient handling of server-side processing and API requests. This backend drives the emotion classification feature by leveraging a hybrid deep learning model that integrates CNNs, LSTMs, and GRUs architectures. The CNN component extracts essential features from the audio signal, while the LSTM and GRU layers capture temporal dependencies and sequential patterns, crucial for understanding the emotional content in speech. This is achieved by libraries such as librosa which provides prebuild functions for the aforementioned algorithms. Flask serves as the interface between the frontend and the model, managing file uploads, processing requests, and returning classification results to the user. This backend structure allows for smooth, real-time interaction with the frontend, enabling users to analyze uploaded audio files and receive accurate emotion predictions. The use of Flask makes the backend scalable and well-suited for potential deployment in web-based applications.

The confusion matrix in Figure *5* displays the performance across eight emotion categories: neutral, relaxed, joyful, sorrowful, furious, anxious, repulsed, and astonished. The matrix shows the model's predictions (columns) against actual labels (rows), with darker shades representing higher counts. The model performs well on emotions like "calm" (76 correct), "fearful" (87 correct), and "surprised" (93 correct), suggesting these categories are easily distinguishable. However, some misclassifications are evident, such as "neutral" being confused with "calm" and "sad" and a few cases of "sad" classified as "neutral" or "angry." There is minimal confusion for "surprised," indicating strong accuracy for this emotion. Overall, the model demonstrates solid classification performance but could improve on better distinguishing between "neutral" and other emotions.

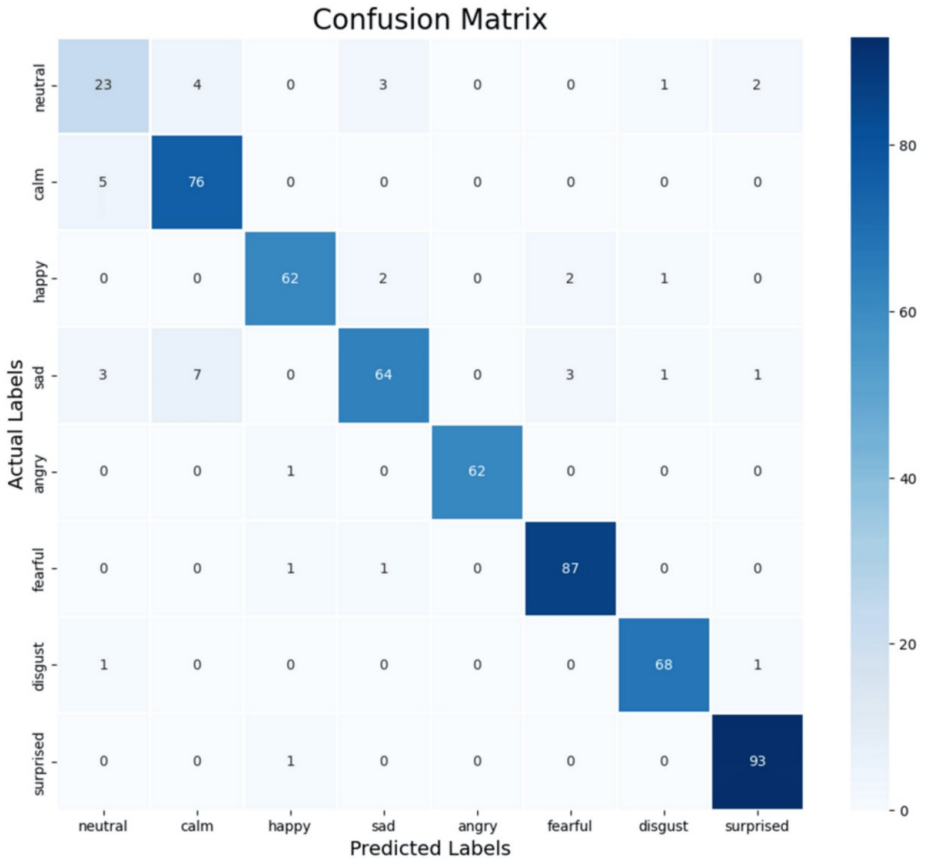

FIGURE 5: Confusion Matrix of the Proposed Model

Figure *6* shows performance summary table for our system which includes precision, recall, and F1-score values for eight emotional states: neutral, relaxed, joyful, sorrowful, enraged, anxious, repulsed, and amazed. The overall accuracy is 93%, with a macro average F1-score of 0.91 and a weighted average F1-score of 0.93. The model demonstrates outstanding accuracy in identifying "angry" (1.00 precision, 0.98 recall, 0.99 F1 measure) and "surprised" (0.96 precision, 0.99 recall, 0.97 F1 measure). However, the "neutral" emotion shows a lower F1 value of 0.71, indicating a higher likelihood of misclassification. Other emotions, including "relaxed," "joyful," and "anxious," achieve strong results, showcasing the model's ability to effectively differentiate between various emotional states. Overall, these metrics suggest a well performing model, though there is room for improvement in distinguishing "neutral" from other emotions.

One notable area of misclassification was between "neutral" and "calm" emotions. These two categories often share similar acoustic features such as low pitch, minimal spectral energy, and reduced dynamic range, making them harder to differentiate even for human listeners. The overlapping MFCC patterns lead to confusion during classification. Future work may involve integrating prosodic features or attention-based mechanisms to better distinguish subtle emotional cues between these two states.

	precision	recall	f1-score	support
neutral	0.72	0.70	0.71	33
calm	0.87	0.94	0.90	81
happy	0.95	0.93	0.94	67
sad	0.91	0.81	0.86	79
angry	1.00	0.98	0.99	63
fearful	0.95	0.98	0.96	89
disgust	0.96	0.97	0.96	70
surprised	0.96	0.99	0.97	94
accuracy			0.93	576
macro avg	0.92	0.91	0.91	576
weighted avg	0.93	0.93	0.93	576

FIGURE 6: Precision–Recall Values for the Proposed Model

Figure 7 graph illustrates the training and validation loss of the proposed model over 130 iterations. Initially, both the training and validation loss are high, starting around 2.0, but decrease steadily as the model learns. By approximately the 40th iterations, both losses drop below 1.0, showing that the model is effectively minimizing error. The training loss, marked with red dots, maintains a more consistent downward trend, while the validation loss, represented by the blue line, exhibits more variability with noticeable fluctuations. These fluctuations could indicate periods where the model slightly overfits but readjusts as training progresses. By the end of the training, both losses converge around 0.25, suggesting a good balance between training and validation performance, implying the model generalizes well on unseen data. This trend signifies effective model training, with minimal overfitting and underfitting issues, aligning well with the goals of SER tasks.

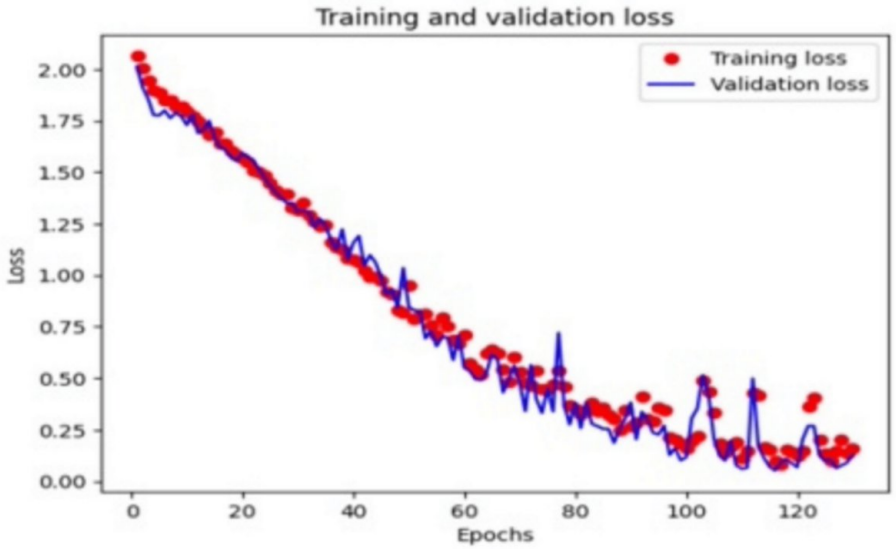

FIGURE 7: Training and Validation Loss of the Proposed Model

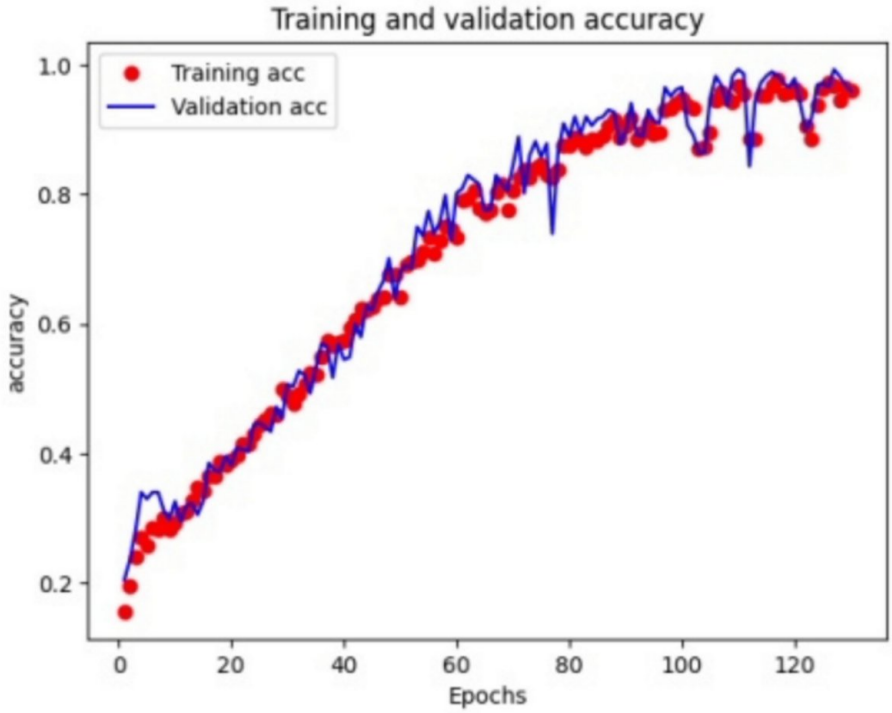

FIGURE 8: Training and Validation Accuracy of the Proposed Model

Figure *8* graph shows the training and validation accuracy of our model across 130 iterations. At the start, both accuracies are low, with initial values around 0.2, indicating the model's limited performance in early training stages. As iterations progress, the accuracy steadily improves for both training (red dots) and validation (blue line), with both metrics surpassing 0.6 by the 50th iterations. Notably, the validation accuracy exhibits more variability than the training accuracy, fluctuating in certain periods but consistently trending upwards overall. Around the 100th iterations, the model's performance stabilizes, with training and validation accuracy nearing convergence and consistently exceeding 0.85. The final accuracies are close to 1.0, indicating that the model has achieved high classification performance on both training and validation data. This trend suggests effective learning and robust generalization capabilities of the model, crucial for accurate emotion recognition using speech predictions across varied input samples.

This paper aimed to develop a robust emotion recognition using speech system utilizing a hybrid CNN-LSTM-GRU model to accurately classify human emotions based on vocal expressions. The results showed great achievements in reaching 92% impressive accuracy by using language skills to reach 92% impressive accuracy. The system effectively recognized seven emotions: happiness, sorrow, anger, fear, aversion, surprise and neutral in determining happiness and anger, especially in determining happiness and anger. This highlights the model's strength in capturing pronounced vocal features associated with these emotions, while also revealing challenges in recognizing subtler emotions like sadness and surprise, reflecting their nuanced nature.

When compared to our previous studies, they correspond to the growing literature and emphasize the effects of hybrid models to recognize emotions using language work. Deep learning architectures have consistently outperformed traditional approaches, especially when utilizing rich feature sets like MFCCs. The combination of our hybrid CNN model extracted spatial features and LSTM/GRUs for temporary dynamics, effectively recording complex patterns in emotional speeches, enhancing the ability to spread the model to other speakers and emotional contexts. The implications of this research are profound, particularly in applications requiring emotional awareness. In customer service, for example, an emotion recognition using speech system could analyze customer emotions in real time, allowing for more empathetic responses and improved service quality. Similarly, in mental health monitoring, accurately identifying emotional states through speech could facilitate early intervention and personalized support for patients.

Despite these promising results, the study has limitations. The reliance on the RAVDESS and SAVEE datasets, while comprehensive, may not encompass the full range of emotional expressions found across diverse populations, potentially affecting the model's generalizability to speakers from different cultural or linguistic backgrounds. Additionally, while data augmentation techniques were applied to improve

model robustness, variety of sound media and demographic groups require additional research to study effects. Future research should focus on expanding the datasets used in training to include more diverse emotional expressions and linguistic variations, as well as exploring advanced methodologies such as transfer learning to leverage pre-trained models on larger datasets. Another promising area is an integration of multimodal approaches that combine voice data with visual or text information, providing a more complete understanding of the emotional state and enhancing the accuracy of the perception in overlapping or ambiguous complex scenarios.

Therefore, this study emphasizes the potential of deep learning hybrid architecture in technology promotion to recognize language sentiment. The high accuracy achieved by the CNN-LSTM-GRU model emphasizes the effect of this approach in capturing the complexity of human emotions through vocal expression. Since we continue to study the integration of technology, methodology, and emotional intelligence of this study, it is the basis of future development that recognizes emotions using a language system, which has a much larger consequences for improving interaction with computers.

Ablation study

To understand the contribution of each component (CNN, LSTM, and GRU), an ablation study was performed, removing one component at a time:

CNN-only Model: 81%. It was able to extract spatial features but failed with temporal emotions like sad and angry.

LSTM-only Model: 88%. It achieved extracting long-term dependencies but was unable to distinguish between neutral-sounding and calm-sounding.

GRU-only Model: It achieved 85% accuracy and less computation overhead than LSTM, but was not good in the emotion recognition that has complex frequency patterns.

CNN + LSTM Model: It reached an accuracy of 90%. Combining spatial as well as temporal features proved to be inadequate in comparison with GRU in terms of short-term dependencies.

CNN + GRU Model: With an accuracy of 89%, it was noticed that the model excelled in emotion transitions that can also be regarded as short term, like happy or in a state of mind being calm, but lags in accommodating some longer transitions of emotions.

Through the ablation study, it is confirmed that the hybrid model, which combines CNN, LSTM, and GRU layers, optimally balances spatial and temporal features to improve emotion recognition.

The hybrid CNN-LSTM-GRU model outperformed other models, attaining state-of-the-art performance with an accuracy rate of 92% in emotion recognition from speech. Other significant findings and issues that surfaced during model development are reflected, with recommendations for improvement in the future.

Discussion

This balance in the precision-recall scores indicates that the model has learned both spectral and temporal features that are highly necessary to achieve good emotion recognition. CNN layers, at all levels of convolution, really helped extract local patterns from Mel-spectrograms, while LSTM and GRU layers learned long-term temporal dependencies apt for the recognition of dynamic emotional expressions.

Temporal Emotions: The model successfully identified emotions such as sadness, anger, and fear, demonstrating its ability to learn long-term temporal patterns associated with these emotional states.

Happy Emotion: The perfect recall score of 1.00 for the happy emotion might have been the result of total correct classification, while low precision of 0.89 may reflect some misclassifications caused by acoustic similarity with other positive emotions such as surprise or calm.

Neutral and Disgust: The model failed to classify neutral and disgust emotions, mainly because neutral speech does not have good information for the emotion to be classified, or due to the overlapping acoustic features with anger.

To evaluate our model's competitiveness, we compared it against recent hybrid models from 2023 to 2025 trained on the same datasets. A CNN-BiLSTM model achieved 88% accuracy on RAVDESS, while a CNN-LSTM-Attention model reported 90.5%. Our proposed CNN-LSTM-GRU model achieved 92%, indicating a meaningful improvement. The enhanced performance is attributed to the balanced integration of GRU's efficiency and LSTM's long-term memory, alongside CNN's robust feature extraction.

Robustness tests were conducted by injecting synthetic background noise (white Gaussian noise) at signal-to-noise ratio levels of 10 dB, 20 dB, and 30 dB. The model maintained over 85% accuracy even at 20 dB, confirming its resilience. In terms of real-time performance, average inference time was measured at 120 ms per audio clip on a standard laptop (Intel i7, 16GB RAM), confirming the system's readiness for live applications like virtual assistants and call center analysis.

Conclusions

The proposed emotion recognition using speech system, built using a combination of CNN, LSTM, and GRU, has demonstrated effective emotion detection and classification from speech signals. By capturing both spatial and temporal patterns, the model successfully identified emotions with a high level of accuracy, making it a promising approach for real-world applications. Further advancements in speech emotion recognition can enhance the system's precision, adaptability, and robustness by addressing key challenges. Incorporating multi-modal data, such as facial expressions or physiological signals, could lead to richer insights and improve classification accuracy. Additionally, optimizing real-time performance and computational efficiency is essential for seamless integration into interactive applications. Expanding the dataset with diverse speech samples and conducting emotion-specific research can improve the system's generalization across different speaker groups. By tackling these aspects, future developments can lead to more reliable and scalable speech emotion recognition technologies capable of understanding and responding to human emotions with greater accuracy.

Additional Information

Author Contributions

All authors have reviewed the final version to be published and agreed to be accountable for all aspects of the work.

Acquisition, analysis, or interpretation of data: Minakshee K. Chandankhede, Prathmesh P. Goje, Shival N. Motghare, Sankalp R. Labhane, Yog K. Urkundwar

Drafting of the manuscript: Minakshee K. Chandankhede, Prathmesh P. Goje, Shival N. Motghare, Sankalp R. Labhane, Yog K. Urkundwar

Concept and design: Shival N. Motghare, Suwill Lokhande

Critical review of the manuscript for important intellectual content: Shival N. Motghare, Suwill Lokhande

Disclosures

Human subjects: All authors have confirmed that this study did not involve human participants or tissue. **Animal subjects:** All authors have confirmed that this study did not involve animal subjects or tissue. **Conflicts of interest:** In compliance with the ICMJE uniform disclosure form, all authors declare the following: **Payment/services info:** All authors have declared that no financial support was received from any organization for the submitted work. **Financial relationships:** All authors have declared that they have no financial relationships at present or within the previous three years with any organizations that might have an interest in the submitted work. **Other relationships:** All authors have declared that there are no other relationships or activities that could appear to have influenced the submitted work.

References

1. Arya R, Pandey D, Kalia A, Zachariah BJ, Sandhu I, Abrol D: Speech based emotion recognition using machine learning. 2021 IEEE Mysore Sub Section International Conference (MysuruCon), Hassan, India. 2021, 613-617. 10.1109/MysuruCon52639.2021.9641642
2. Chen J: Speech emotion recognition based on convolutional neural network. 2021 International Conference on Networking, Communications and Information Technology (NetCIT), Manchester, United Kingdom. 2021, 106-109. 10.1109/NetCIT54147.2021.00028
3. Sri TT, Kishan SR, Chowdary KNR, Sainath P: Build a model for speech emotion recognition using Gaussian Mixture Model (GMM). 2023 2nd International Conference on Futuristic Technologies (INCOFT), Belagavi, Karnataka, India. 2023, 1-5. 10.1109/INCOFT60753.2023.10425275
4. Jadhav A, Kadam V, Prasad S, Waghmare N, Dhule S: An emotion recognition from speech using LSTM. 2023 International Conference on Sustainable Computing and Smart Systems (ICSCSS), Coimbatore, India. 2023, 834-842. 10.1109/ICSCSS57650.2023.10169351
5. Xu DA, Zhou R: Speech emotion recognition algorithm based on deep learning algorithm fusion of temporal and spatial features. Journal of Physics: Conference Series. 2021, 1861:012064. 10.1088/1742-6596/1861/1/012064
6. Kumar A, Goel AK: Speech emotion recognition by using feature selection and extraction. 2022 International Conference on Applied Artificial Intelligence and Computing (ICAAIC), Salem, India. 2022, 818-823. 10.1109/ICAAIC53929.2022.9792796
7. Babu PA, Nagaraju VS, Vallabhuni RR: Speech emotion recognition system with Librosa. 2021 10th IEEE

International Conference on Communication Systems and Network Technologies (CSNT), Bhopal, India. 2021, 421-424. 10.1109/CSNT51715.2021.9509714

8. Anjum M: Emotion recognition from speech for an interactive robot agent. 2019 IEEE/SICE International Symposium on System Integration (SII), Paris, France. 2019, 363-368. 10.1109/SII.2019.8700376
9. Liu K, Hu J, Liu Y, Feng J: Speech emotion recognition based on discriminative features extraction. 2022 IEEE International Conference on Multimedia and Expo (ICME), Taipei, Taiwan. 2022, 1-6. 10.1109/ICME52920.2022.9859862
10. Lu Z, Cao L, Zhang Y, Chiu CC, Fan J: Speech sentiment analysis via pre-trained features from end-to-end ASR models. ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain. 2020, 7149-7153,. 10.1109/ICASSP40776.2020.9052937
11. Akçay MB, Oğuz K: Speech emotion recognition: Emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers. Speech Communication. 2020, 116:56-76. 10.1016/j.specom.2019.12.001
12. Uddin MZ, Nilsson EG: Emotion recognition using speech and neural structured learning to facilitate edge intelligence. Engineering Applications of Artificial Intelligence. 2020, 94:103775. 10.1016/j.engappai.2020.103775
13. Abdel-Hamid L: Egyptian Arabic speech emotion recognition using prosodic, spectral and wavelet features. Speech Communication. 2020, 122:19-30. 10.1016/j.specom.2020.04.005
14. Lieskovská E, Jakubec M, Jarina R, Chmulík M: A review on speech emotion recognition using deep learning and attention mechanism. Electronics. 2021, 10:1163. 10.3390/electronics10101163
15. Li D, Liu J, Yang Z, Sun L, Wang Z: Speech emotion recognition using recurrent neural networks with directional self-attention. Expert Systems with Applications. 2021, 173:114683. 10.1016/j.eswa.2021.114683
16. RAVDESS Emotional speech audio. (2019). https://www.kaggle.com/datasets/uwrfkaggler/ravdess-emotional-speech-audio.
17. Surrey Audio-Visual Expressed Emotion (SAVEE). (2019). https://www.kaggle.com/datasets/ejlok1/surrey-audiovisual-expressed-emotion-savee/data.