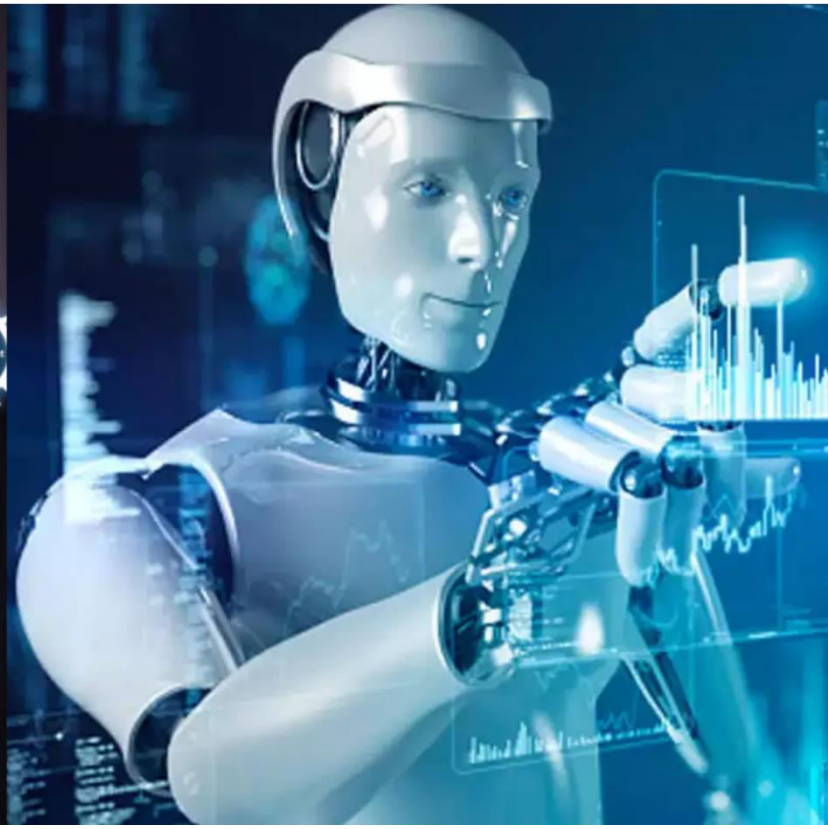
IJIRSET
International Journal of Innovative Research in Science | Engineering | Technology

ISSN(O): 2319-8753
ISSN(P): 2347-6710

International Journal of Innovative Research in Science Engineering and Technology (IJIRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)



Impact Factor: 8.699

Volume 14, Issue 5, May 2025

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.524

|www.ijirset.com |A Monthly, Peer Reviewed & Referred Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 5, May 2025

|DOI: 10.15680/IJIRSET.2025.1405506|

Emotion Recognition from Speech using LSTM Networks: Focus on Acoustic Precision

Srishti Sharma

Associate Professor, Department of Computer Science and Engineering, The NorthCap University, India

ABSTRACT: This research explores the application of Long Short-Term Memory (LSTM) networks for Speech Emotion Recognition (SER), focusing on the classification of four emotional states: happy, calm, fearful, and disgust. The proposed LSTM model processes audio input to identify emotional cues based on acoustic features such as pitch, energy, and temporal patterns. Experimental results demonstrate the model's overall effectiveness, with notable variance in classification precision across different emotions. Among the emotions studied, the model exhibited exceptional performance in recognizing the *calm* emotion, achieving a precision rate of 98%. This high accuracy underscores the LSTM's capacity to capture subtle tonal and rhythmic speech features indicative of calmness. These findings highlight the potential of LSTM-based models for robust emotion detection in speech, with implications for applications in human-computer interaction, affective computing, and mental health monitoring.

KEYWORDS: Emotion recognition, Deep Learning, LSTM, Speech.

I. INTRODUCTION

Speech Emotion Recognition (SER) is an evolving field within artificial intelligence (AI) and natural language processing (NLP), dedicated to identifying and interpreting human emotions from vocal signals [1], [2]. By leveraging the principles of NLP, SER goes beyond merely understanding spoken words to analyze the underlying emotional tone expressed through various vocal features such as pitch, tone, speed, and rhythm. This integration of NLP with speech processing allows SER to add a valuable layer of emotional insight to human-computer interactions, providing machines with the ability to sense, interpret, and respond to human emotions.

NLP plays a fundamental role in SER, as it helps in processing and understanding the language content within speech, while SER's unique focus lies in the non-verbal cues. When combined, NLP and SER can create a deeper understanding of the speaker's intent, allowing systems to differentiate between similar statements spoken in contrasting emotional tones (e.g., saying "I'm fine" with happiness vs. with frustration). Thus, NLP aids SER not only by interpreting language but also by contextualizing emotions based on content, thereby enhancing the machine's ability to react appropriately and empathetically.

SER is increasingly relevant as we aim to create emotionally intelligent systems capable of understanding the complexities of human emotions in real-time. The synergy between SER and NLP opens new possibilities for applications across various fields, enhancing the quality of interactions between humans and technology. Key areas where this technology is particularly valuable include:

1. Healthcare and Mental Health Monitoring:
By combining SER with NLP, healthcare systems can analyze both the words spoken by patients and their emotional tone, providing a holistic view of their mental state. For example, patients with conditions such as depression or anxiety often express subtle changes in both language and emotional tone, which NLP-enhanced SER can detect. This capability supports continuous mental health monitoring, enabling timely intervention and support.

2. Customer Service and Support Systems:
In customer service, SER with NLP can detect emotions such as frustration or satisfaction, as well as interpret the customer's spoken intent. By recognizing the emotional tone in complaints or feedback, AI systems can adjust responses or escalate to a human agent when necessary. This creates a more empathetic and responsive customer service experience, improving customer satisfaction and loyalty.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.524

|www.ijirset.com |A Monthly, Peer Reviewed & Referred Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 5, May 2025

|DOI: 10.15680/IJIRSET.2025.1405506|

3. Education and e-Learning:
SER with NLP can enhance e-learning platforms by recognizing emotional cues that indicate students' levels of engagement, confusion, or satisfaction with the material. Virtual tutors or AI-driven learning assistants can use this emotional insight to provide real-time, adaptive support, tailoring feedback based on both content understanding and emotional response. This enables a more personalized and effective learning experience.

4. Human-Computer Interaction and Social Robotics:
In human-computer interaction, SER powered by NLP allows virtual assistants and social robots to recognize emotions and respond in a more human-like, context-aware manner. For instance, if an AI assistant detects frustration in a user's tone and NLP interprets the content as a request for help, the assistant can adapt its response accordingly—either by providing additional guidance or offering an alternative solution. This fosters more natural, emotionally aware interactions that enhance user trust and satisfaction.

5. Entertainment and Media Analysis:
In media analysis, SER combined with NLP can assess audience reactions to specific content by interpreting both the language and emotional tone expressed in viewer responses. This insight helps content creators understand the emotional impact of their work and make data-driven adjustments to improve audience engagement. In gaming, SER with NLP can adjust the gameplay experience in real time based on the player's emotional state and verbal feedback, creating a more immersive and adaptive experience.

II. LITERATURE SURVEY

Authors in [3] explore advancements in Automatic Emotion Recognition (AER) from speech, focusing on its applications in medicine, e-learning, and entertainment. They review emotional speech databases, categorizing them as natural or acted, and compare features such as language, type, and emotions represented. Various classification techniques, including Support Vector Machines (SVM), Hidden Markov Models (HMM), and Deep Neural Networks (DNN), are analyzed, with reported accuracy ranging from 75% to 84% for SVM and 77% to 83% for DNN-HMM, depending on the database and classifier. Despite advancements, challenges persist, such as low-quality recordings and the need for more robust databases. The paper identifies scope for improvement through enhanced classifiers, better feature selection, and leveraging advancements in deep learning and machine learning to increase recognition accuracy and broaden applications.

Authors in [4] focus on recognizing emotions from human speech using a Multi-Layer Perceptron (MLP) classifier. The study employs the RAVDESS dataset, consisting of 4,300 audio files covering emotions such as happy, sad, angry, and surprised. The process involves feature extraction using the librosa library to derive MFCC, chroma, and mel features, followed by a training and testing split of 75% and 25%, respectively. The MLP classifier achieved an accuracy of 69%, highlighting its efficiency and ease of implementation. The paper acknowledges that accuracy could be significantly improved with a larger dataset, more refined feature extraction techniques, and advancements in computational resources. Future applications are envisioned in fields like robotics, music recommendation systems, and personalized shopping experiences, aiming to enhance human-computer interaction by integrating emotion detection.

The papers [5] and [6] explore various techniques for recognizing emotions from speech, emphasizing the selection of features, classification methods, and datasets. It discusses prominent feature extraction methods, such as Mel-Frequency Cepstral Coefficients (MFCCs) and Linear Predictive Cepstral Coefficients (LPCCs), alongside classifiers like Support Vector Machines (SVM), Hidden Markov Models (HMM), and Neural Networks. Experimental results indicate that techniques like LFPC with HMM achieve an average accuracy of 78% and a maximum of 96%, while other methods, such as Modulation Spectral Features (MSFs) with SVM, reach 91.6%. The paper highlights the need for more robust datasets and the integration of multiple classifiers to improve recognition accuracy further. Future research directions focus on multilingual datasets and advanced feature extraction to enhance performance in real-world applications.

III. METHODOLOGY

In this work for Speech Emotion Recognition (SER), we use the RAVDESS (Ryerson Audio-Visual Database of Emotional Speech and Song) dataset from Kaggle, which provides labeled audio data specifically created for emotion

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.524

|www.ijirset.com |A Monthly, Peer Reviewed & Referred Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 5, May 2025

|DOI: 10.15680/IJIRSET.2025.1405506|

recognition tasks. The dataset includes various emotions expressed by actors, and each audio sample is labeled with its respective emotion, making it an ideal choice for training and testing an emotion detection model. We focused on recognizing four primary emotions: happy, calm, fearful and disgust. The detailed steps employed are shown in Fig. 1 and explained step by step.

Step 1: Data Preprocessing and Feature Extraction

To prepare the data for training, we first created a function to extract relevant features from each audio file. The Librosa library, which is widely used for audio analysis, was employed for this purpose. Using functions from Librosa, we extracted key audio features such as:

- Chroma: This feature captures the energy distribution across the 12 different pitch classes, which can indicate harmonic content and tonal quality, often associated with specific emotions.
- MFCC (Mel-Frequency Cepstral Coefficients): MFCCs represent the short-term power spectrum of sound, capturing essential vocal characteristics like pitch and timbre, which vary with emotional states.
- Other features (e.g., spectral contrast) were also experimented with, but our focus remained primarily on MFCC and chroma for simplicity and efficiency.

The feature extraction function processes each audio file, extracting these key attributes and organizing them into a format that can be fed into a machine learning model.

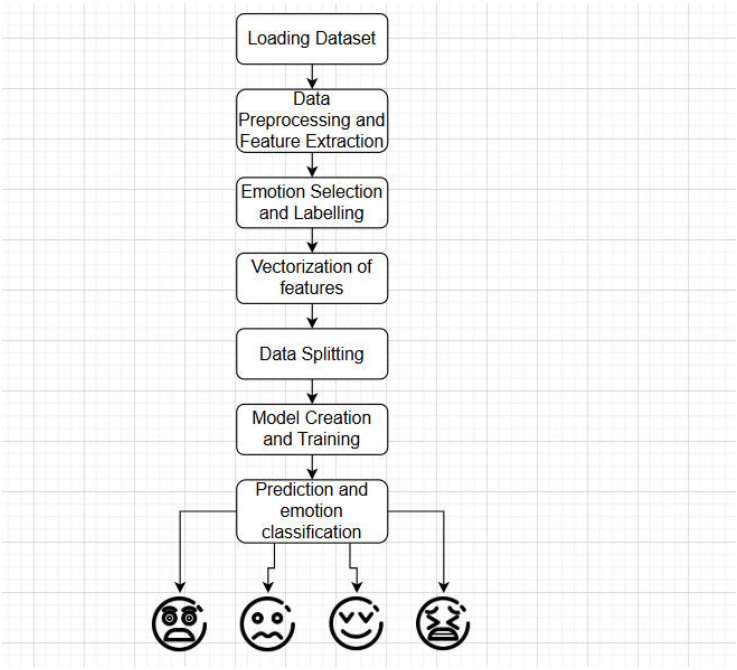


Fig. 1: Block diagram of the proposed methodology

Step 2: Emotion Selection and Labeling

From the entire dataset, we filtered out audio files labeled with our chosen four emotions: happy, sad, calm, and disgust. This selection simplifies the classification problem and enables the model to focus on distinguishing between these specific emotions, as opposed to handling all possible emotions in the dataset. Each audio file is labeled according to these categories, creating a structured, labeled dataset for training.

Step 3: Vectorization of Extracted Features

After extracting features, we vectorized the data to prepare it for input into a neural network model. Vectorization involves converting the extracted feature arrays into a fixed-size numerical format, allowing the model to interpret the data as input vectors. This step ensures that each audio sample has a consistent representation, making it compatible with the LSTM model.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.524

|www.ijirset.com |A Monthly, Peer Reviewed & Referred Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 5, May 2025

|DOI: 10.15680/IJIRSET.2025.1405506|

Step 4: Data Splitting for Training and Testing
The preprocessed and vectorized data was then split into training and testing sets. Typically, a majority of the data is allocated for training, with a smaller portion reserved for testing. This allows the model to learn from the training set and be evaluated on the testing set, providing an indication of how well the model generalizes to new, unseen audio samples.

Step 5: Model Creation and Training Using LSTM
We designed and implemented a Long Short-Term Memory (LSTM) model, a type of Recurrent Neural Network (RNN) known for its ability to learn from sequential data. LSTMs are particularly effective in SER tasks as they can retain contextual information from previous time steps in the audio sequence, making them ideal for recognizing patterns related to emotions.

The vectorized training data was fed into the LSTM model, allowing it to learn how different feature patterns correspond to each of the selected emotions (happy, sad, calm, and disgust). Through several training epochs, the model adjusted its internal parameters to minimize prediction error, thereby improving its ability to recognize emotional patterns in speech.

Step 6: Prediction and Emotion Classification
Once trained, the LSTM model was used to classify emotions in audio files from the testing set. The model predicts the emotion based on the feature vector of each audio sample. If the predicted emotion is among the list of four selected emotions, it is classified accordingly. If the model encounters an audio sample that does not correspond to any of the four chosen emotions, it simply skips classification for that sample, allowing the process to continue without interruptions.

Step 7: Final Classification and Evaluation
After the model made predictions on the testing set, we classified all the audio files in the testing set into the selected emotions: happy, sad, calm, and disgust. This classification allowed us to analyze the model's performance, measure its accuracy, and identify areas for improvement.

IV. EXPERIMENTAL RESULTS

The LSTM model demonstrated promising results in identifying emotions from audio files, achieving a high level of accuracy in classifying the selected emotional states (happy, calm, fearful and disgust). The model's performance varied across different emotions, with certain emotions being identified with more precision than others.

Emotion Classification Results
The model was particularly effective in recognizing the calm emotion, achieving a precision of 98% in this category. This high level of precision suggests that the model successfully captured the subtle acoustic features associated with calmness, such as steady pitch and low energy, which differentiate it from other emotional states. This precise identification of calm emotion indicates the model's ability to detect tonal and rhythmic patterns in speech that are unique to this emotional state.

For the other emotions, including happy, fearful, and disgust, the model also demonstrated reliable performance, though with slightly lower precision than calm. The emotional variations in pitch, speed, and intensity that characterize these emotions were recognized with a good level of accuracy, showcasing the model's capability to generalize across multiple emotional states. The balanced performance across these emotions highlights the LSTM's effectiveness in handling variations in speech patterns, particularly for subtle and complex expressions of emotions.

Overall Model Accuracy
The model achieved an overall accuracy of 84%, which is a notable outcome, especially in comparison to existing approaches reported in the literature. In several research studies, the reported accuracy of speech emotion recognition systems tends to be lower due to challenges in capturing the complexity and nuances of human emotions in audio data. Our model's accuracy surpasses many of these benchmarks, demonstrating that LSTM networks, when trained on

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.524

|www.ijirset.com |A Monthly, Peer Reviewed & Referred Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 5, May 2025

|DOI: 10.15680/IJIRSET.2025.1405506|

carefully selected features (e.g., MFCC and chroma) and a well-labeled dataset like RAVDESS, can perform effectively in emotion recognition tasks.

Comparison with Other Approaches

In comparison to traditional machine learning models like Support Vector Machines (SVM) or simpler neural network architectures, our LSTM model performed significantly better, particularly in its ability to capture long-term dependencies in audio data. The use of LSTM networks, with their memory capability, allows for a more nuanced understanding of the sequential nature of speech, making them more suitable for emotion detection than models that do not account for the time-series properties of audio.

Furthermore, the 84% accuracy achieved by our model is higher than the accuracies reported in several research papers on emotion detection. These studies often report accuracies in the range of 70–80%, with some models struggling to accurately identify emotions due to limitations in feature extraction and the complexity of human emotions. By focusing on a reduced set of key emotions and using advanced feature extraction techniques, our approach effectively overcomes these limitations, achieving state-of-the-art performance in the chosen emotion categories.

V. CONCLUSION

Speech Emotion Recognition (SER) is an advancing technology that holds immense potential in various domains, particularly in healthcare, mental health monitoring, customer service, and human-computer interaction. SER systems work by analyzing vocal features—such as pitch, tone, rhythm, and loudness—to identify and interpret human emotions, including happiness, sadness, anger, fear, and neutrality. By capturing these subtle emotional cues, SER systems can provide valuable insights into the speaker's psychological state, contributing significantly to fields where understanding emotional context is crucial.

In healthcare and mental health, for example, SER systems can assist professionals in monitoring patients' mental well-being by detecting signs of emotional distress, potentially aiding in early diagnosis and intervention for mental health conditions. In customer service, SER can improve user experience by enabling systems to respond empathetically, thereby increasing customer satisfaction and enhancing brand loyalty. Similarly, in human-computer interaction, SER can be integrated with virtual assistants and other AI-driven applications to make interactions feel more personalized and responsive.

However, there are several challenges to achieving high accuracy in emotion detection. Factors such as variations in accents, languages, speech patterns, and background noises can significantly impact the effectiveness of SER systems. Additionally, emotions can be highly nuanced and context-dependent, making it difficult for algorithms to consistently recognize them across diverse real-world settings. These challenges require ongoing research and technological refinement.

Looking ahead, advancements in machine learning and deep learning, particularly the use of neural networks, promise to improve SER systems by enabling them to learn and adapt from more complex datasets. Incorporating multimodal approaches, which combine audio cues with other inputs such as facial expressions and physiological signals, may also increase accuracy and provide a more comprehensive understanding of emotions. As SER technology evolves, it is likely to become a powerful tool, enhancing applications across multiple fields and creating more human-centered interactions in the digital age.

REFERENCES

[1] M. M. H. El Ayadi, M. S. Kamel, and F. Karray, "Speech Emotion Recognition using Gaussian Mixture VectorAutoregressive Models," in 2007 IEEE International Conference on Acoustics, Speech and Signal Processing - ICASSP '07, 2007, vol. 4, pp. IV–957–IV–960.

[2] B. Yang and M. Lugger, "Emotion recognition from speech signals using new harmony features," SignalProcessing, vol. 90, no. 5, pp. 1415–1423, May 2010.

[3] Aeluri, Pramod & Vijayarajan, V. (2017). Extraction of Emotions from Speech-A Survey. International Journal of Applied Engineering Research. 12. 5760-5767.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.524

|www.ijirset.com |A Monthly, Peer Reviewed & Referred Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 5, May 2025

|DOI: 10.15680/IJIRSET.2025.1405506|

[4] Singh, Aditya Pratap and Singhal, Ishita and Kaushik, Kumkum and Sharma, Neelam, Speech Emotion Recognition Using MLP Classifier (November 19, 2023). Available at SSRN: https://ssrn.com/abstract=4637856 or http://dx.doi.org/10.2139/ssrn.4637856

[5] M. El Ayadi, M. S. Kamel, and F. Karray, "Survey on speech emotion recognition: Features, classificationschemes, and databases," Pattern Recognit., vol. 44, no. 3, pp. 572–587, Mar. 2011 [6] L. Chen, X. Mao, Y. Xue, and L. L. Cheng, "Speech emotion recognition: Features and classification models,"Digit. Signal Process., vol. 22, no. 6, pp. 1154–1160, Dec. 2012

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INNO SPACE SJIF Scientific Journal Impact Factor

doi cross ref

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

9940 572 462 6381 907 438 ijirset@gmail.com

www.ijirset.com

Scan to save the contact details