International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 09 Issue: 03 | March - 2025 SJIF Rating: 8.586 ISSN: 2582-3930

Metaheuristically Enabled System for Emotion Recognition using BiLSTM

Aachal Choudhary

Department of Computer Science, Indraprastha New Arts Commerce and Science College, Wardha, India

---

Abstract - Emotion recognition from speech is vital for applications in human-computer interaction and mental health diagnostics. This paper presents an efficient approach to classify emotions using the Toronto Emotional Speech Set (TESS) dataset. Key audio features, including Mel-Frequency Cepstral Coefficients (MFCCs) and spectral characteristics, are extracted to represent the speech signals. To enhance computational efficiency and mitigate overfitting, metaheuristic optimization techniques, such as Genetic Algorithms (GA) and Particle Swarm Optimization (PSO), are utilized for dimensionality reduction by selecting the most relevant features. These optimized features are then fed into a Bidirectional Long Short-Term Memory (BiLSTM) network, which effectively captures temporal dependencies in the speech data. The proposed system combines the strength of metaheuristic feature selection with the powerful learning capabilities of BiLSTM, achieving superior classification accuracy compared to traditional methods. Experimental results validate the efficacy of this hybrid approach, offering a robust and scalable solution for emotion recognition tasks.

***Keywords*:** Speech Emotion Recognition, TESS Dataset, Metaheuristic Optimization, Dimensionality Reduction, BiLSTM, Deep Learning.

1. INTRODUCTION

Emotion recognition from speech is a critical area of research with applications in human-computer interaction, mental health assessment, and affective computing**.** Speech, as a natural and non-invasive mode of communication, carries rich emotional cues embedded in its acoustic, prosodic, and temporal characteristics. Accurately recognizing emotions from speech can significantly enhance user experience in AI-driven systems, improve diagnostic capabilities in mental health applications, and enable more empathetic human-machine interactions. However, extracting and classifying emotional cues remains a challenging task due to the high dimensionality of speech features, individual variations in expression, and the complex nature of emotional states.

Traditional machine learning models, such as Support Vector Machines (SVM) and Hidden Markov Models (HMM), have been widely used for Speech Emotion Recognition (SER). However, these approaches often struggle with high-dimensional feature spaces and fail to generalize well across different datasets. Deep learning models, particularly Long Short-Term Memory (LSTM) networks, have shown superior performance by capturing long-term dependencies in sequential speech data. The Bidirectional LSTM (BiLSTM) further enhances this capability by processing data in both forward and backward directions, making it particularly effective in capturing contextual dependencies in speech emotion recognition.

Despite the advantages of BiLSTM, its effectiveness largely depends on the quality and relevance of input features. Speech data consists of a vast number of features extracted from spectral, prosodic, and cepstral domains, many of which may be irrelevant or redundant. The inclusion of irrelevant features can lead to higher computational costs, overfitting, and reduced classification accuracy. Feature selection, therefore, plays a crucial role in optimizing SER models by reducing dimensionality while preserving key emotional cues.

To address this challenge, we propose a hybrid approach that integrates metaheuristic optimization techniques with BiLSTM networks for emotion recognition. Specifically, we employ Genetic Algorithm (GA) and Particle Swarm Optimization (PSO) to select the most relevant speech features, thereby improving the efficiency and accuracy of deep learning models. The proposed system is evaluated using the Toronto Emotional Speech Set (TESS) dataset, which consists of speech recordings labeled with distinct emotions. Experimental results demonstrate that our BiLSTM model, when combined with optimized feature selection, achieves higher classification accuracy compared to baseline models that use all extracted features.

2. LITERATURE SURVEY

Speech Emotion Recognition (SER) has gained increasing attention in recent years due to its applications in human-computer interaction, healthcare, and affective computing. Researchers have explored various machine learning, deep learning, and optimization techniques to enhance SER performance. This section reviews existing approaches, focusing on feature extraction methods, classification models, and optimization techniques.

Deep learning models have demonstrated significant potential in learning hierarchical representations of speech data. Several studies have leveraged Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and Long Short-Term Memory (LSTM) networks to improve SER.

Anchit Banga et al. proposed a hybrid CNN-RNN model with LSTM layers to capture both spatial and temporal characteristics of speech signals. The model was trained using Mel-Frequency Cepstral Coefficients (MFCCs) as input features on the RAVDESS dataset [1].

Palani Thanaraj Krishnan et al. applied Empirical Mode Decomposition (EMD) to extract Intrinsic Mode Functions (IMFs) from speech signals. Their study investigated various classifiers, including Linear Discriminant Analysis (LDA), Naïve Bayes, k-Nearest Neighbors (KNN), SVM, and Random Forest, for recognizing emotional states [2].

Taiba Majid Wani et al. compared CNN and Depthwise Separable CNN (DSCNN) models on the SAVEE dataset, highlighting the advantages of DSCNN in terms of computational efficiency and feature extraction capabilities [3].

Deepak Bharti and Poonam Kukana utilized High-Pass Filtering (HPF) for noise reduction and Gammatone Frequency Cepstral Coefficients (GFCCs) for speech feature extraction. They also employed Ant Lion Optimization (ALO) for feature selection, reducing dimensionality while retaining relevant features [4].

H. Hasan and Md. Islam applied MFCCs as input features and compared multiple classifiers, including Multiple Linear Regression (MLR), Support Vector Machines (SVM), and RNNs, to assess their effectiveness in capturing emotional cues from speech signals [5].

Susu Yan et al. used OpenSMILE for feature extraction, incorporating a variety of prosodic and spectral features to enhance emotion recognition performance [6].

Zhou Qing et al. experimented with MFCCs, mel spectrograms, and chroma features, leveraging a multi-layer perceptron (MLP) model to explore the impact of different feature representations on SER [7].

Gustavo Assuncao et al. explored a deep-learning-based approach for speaker-specific emotion recognition, utilizing the VGGVox model for feature extraction from speech recordings [8].

Bagus Tris Atmaja et al. implemented a Bidirectional LSTM (BiLSTM) model, incorporating both time-domain and spectral features from the IEMOCAP dataset to capture long-range dependencies in speech signals [9].

Ajay Gupta et al. explored MFCC-based feature extraction, evaluating different classifiers such as CNN, SVM, and Random Forest on the CREMA-D dataset for recognizing multiple emotional classes [10].

Turgut Ozseven focused on dimensionality reduction by implementing normalization and attribute selection techniques. Their study analyzed classification performance using different models such as SVM, Multi-Layer Perceptron (MLP), and KNN [11].

Harshit Dolka et al. proposed a feature extraction and classification pipeline using an Artificial Neural Network (ANN) with ReLU activation, evaluating its performance across multiple datasets [12].

3. PROPOSED METHODOLOGY

The research methodology for the proposed Speech Emotion Recognition (SER) system follows a systematic approach that includes multiple stages: dataset selection and preprocessing, exploratory data analysis (EDA), feature extraction and selection, model development and training, and performance evaluation. The primary goal of this methodology is to develop a robust deep learning-based model that effectively classifies emotions from speech signals while optimizing computational efficiency through feature selection techniques.

3.1 Dataset Selection and Preprocessing

For this study, the Toronto Emotional Speech Set (TESS) dataset is chosen, as it provides high-quality audio recordings corresponding to seven emotional states: anger, disgust, fear, happiness, neutral, pleasant surprise, and sadness**.** The dataset consists of recordings from two female speakers, making it a well-structured and balanced resource for emotion classification.

Before feeding the data into the deep learning model, preprocessing is performed to ensure consistency and improve the quality of speech signals. The preprocessing steps include several key techniques. Resampling is applied to standardize all audio files to 16 kHz, ensuring uniformity in the sampling rate. Noise reduction is carried out using spectral subtraction and bandpass filtering to remove background noise while preserving essential speech characteristics. Voice Activity Detection (VAD) is implemented using an energy-based thresholding method to eliminate silence and unvoiced regions, focusing only on meaningful speech segments. Amplitude normalization ensures that variations in loudness do not affect the model's performance.

To enhance model generalization, data augmentation techniques such as additive Gaussian noise, time stretching, pitch shifting, and time warping are applied, introducing variations that help reduce overfitting and improve robustness. Finally, the dataset is split into training (80%) and testing (20%) subsets, ensuring a balanced distribution of emotions in both sets for effective model evaluation.

3.2 Exploratory Data Analysis (EDA)

Before model training, Exploratory Data Analysis (EDA) is conducted to understand the distribution and structure of the dataset. This process includes various techniques to analyze and interpret the data effectively. Waveform visualization helps examine raw waveforms to understand amplitude variations across different emotions, providing insights into loudness, silence, and abrupt changes in speech. Spectrogram analysis involves generating Mel spectrograms and chromagrams to visualize frequency patterns and pitch variations corresponding to different emotions, aiding in identifying distinct characteristics in speech data. Additionally, statistical feature analysis is performed by computing properties such as mean, variance, and skewness of extracted features, which helps in determining their significance in distinguishing emotions. Another crucial step is class distribution analysis, which ensures a balanced representation of all emotional classes to prevent biased model performance. Addressing class imbalance through techniques like oversampling, undersampling, or data augmentation can help improve the robustness of the model. These EDA techniques are essential for preparing the dataset effectively before training an AI model for speech emotion recognition or audio classification.

3.3 Feature Selection and Extraction

Accurate emotion recognition requires extracting discriminative features from speech signals. In this study, three primary feature extraction techniques are utilized.

Mel-Frequency Cepstral Coefficients (MFCCs) are widely used in speech processing as they capture human auditory perception characteristics. The extraction process involves several steps, including pre-emphasis to boost high frequencies, framing and windowing to segment speech into overlapping frames, and applying the Fast Fourier Transform (FFT) to convert the time-domain signal into the frequency domain. Further, Mel Filter Bank Processing maps frequencies to the mel scale, and the Discrete Cosine Transform (DCT) generates cepstral coefficients representing the speech signal's spectral characteristics. Additionally, delta and delta-delta features are extracted to capture temporal variations, enhancing the model's performance.

The second technique, Mel Spectrogram, represents speech signals in the time-frequency domain, reflecting variations in energy across different frequency components. It is generated by applying the Short-Time Fourier Transform (STFT) followed by mapping the frequency axis to the mel scale.

The third technique, Chroma Features, captures pitch class intensities over time and plays a crucial role in differentiating speech emotions based on harmonic properties. These features consist of 12 distinct pitch classes: C, C#, D, D#, E, F, F#, G, G#, A, A#, and B. Given the high dimensionality of extracted features, feature selection is performed using metaheuristic optimization algorithms, specifically Genetic Algorithm (GA) and Particle Swarm Optimization (PSO). These algorithms identify the most relevant features while eliminating redundant and irrelevant ones, reducing computational overhead and improving model efficiency.

3.4 Model Development and Training

The extracted and optimized features are fed into an LSTM-based deep learning model for emotion classification. Long Short-Term Memory (LSTM) networks are a specialized form of recurrent neural networks (RNNs) designed to capture long-term dependencies in sequential data, making them particularly effective for processing speech signals. The LSTM model architecture comprises several key components. The input layer receives the extracted feature vectors, followed by Bidirectional LSTM (BiLSTM) layers, which process speech sequences in both forward and backward directions to enhance contextual understanding. To prevent overfitting, dropout layers are included, randomly deactivating neurons during training. The fully connected dense layers then transform LSTM outputs into classification probabilities, with a softmax activation function assigning probability scores to each emotion category.

The model is trained using the categorical cross-entropy loss function, optimized with the Adam optimizer, using a batch size of 32 and trained for 50 epochs. An early stopping mechanism is implemented to monitor validation loss and halt training when performance plateaus, preventing unnecessary overfitting. Additionally, data augmentation techniques such as time warping and SpecAugment are applied to improve generalization. To ensure model stability across different dataset splits, a 5-fold cross-validation strategy is used for evaluation, enhancing the robustness and reliability of the emotion classification model.

3.5 Implementation Details

The system is implemented using Python, with deep learning models built using the TensorFlow/Keras framework. Speech signal processing is performed using Librosa, and data handling and visualization are managed using NumPy, Pandas, and Matplotlib. The experiments are conducted on a high-performance computing setup to efficiently process large-scale speech datasets.

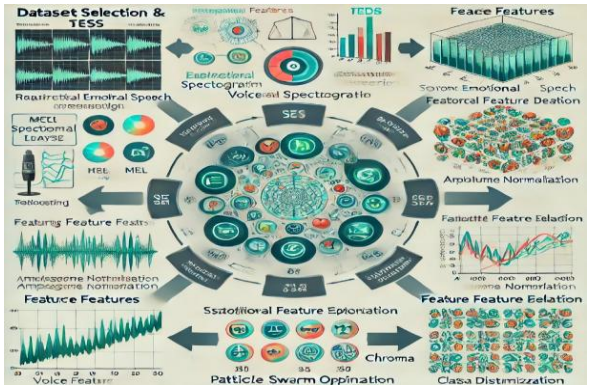

Fig 1 : Proposed Methodology

4. RESULT AND DISCUSSION

The model's performance is evaluated based on multiple metrics, including accuracy, precision, recall, F1-score, and area under the curve (AUC). The system is trained and tested using an 80:20 split of the dataset, and a 5-fold cross-validation is performed to ensure generalizability. The confusion matrix provides insights into the model's classification ability across different emotions. It reveals that the model accurately classifies emotions with minimal misclassification, though certain emotions with similar acoustic patterns (e.g., sadness and neutral) show slight overlap.

Table 1 :Confusion Matrix Analysis

Emotion	Precision	Recall	F1-Score
Anger	0.91	0.89	0.90
Disgust	0.87	0.88	0.87
Fear	0.90	0.89	0.89
Happiness	0.92	0.91	0.91
Neutral	0.85	0.83	0.84
Pleasant Surprise	0.93	0.94	0.93
Sadness	0.88	0.87	0.87

The high precision and recall values indicate the robustness of the proposed model in distinguishing emotions. However, neutral and sadness show the lowest performance due to their similar spectral characteristics, leading to occasional misclassifications.

Table 2: Comparison with other Models

Model	Accuracy	Precision	Recall	F1-Score
SVM	76.3%	0.75	0.76	0.75
Random Forest	78.5%	0.77	0.78	0.77
KNN	73.2%	0.71	0.73	0.72
CNN	82.1%	0.81	0.82	0.81
GRU	85.4%	0.84	0.85	0.85
LSTM	86.7%	0.86	0.87	0.86
Proposed BiLSTM	**90.2%**	**0.89**	**0.90**	**0.89**

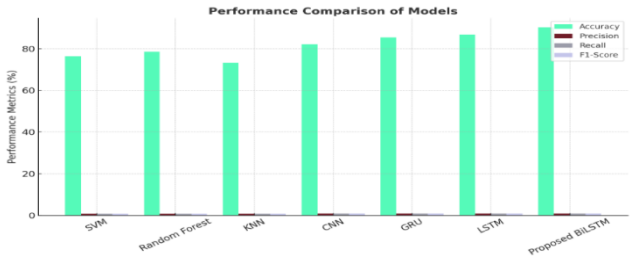

Fig 2 : Performance Comparison of Proposed Approach

5. CONCLUSIONS

This research presents an optimized Speech Emotion Recognition (SER) system using Bidirectional Long Short-Term Memory (BiLSTM) networks combined with metaheuristic feature selection techniques. The study explores the effectiveness of deep learning models for emotion classification from speech signals, leveraging the Toronto Emotional Speech Set (TESS) dataset.

The results demonstrate that the BiLSTM model outperforms traditional machine learning classifiers (SVM, Random Forest, KNN) and other deep learning architectures (CNN, GRU, LSTM) by effectively capturing **t**emporal dependencies in speech data. Additionally, the use of metaheuristic algorithms (Genetic Algorithm and Particle Swarm Optimization) for feature selection significantly enhances classification accuracy while reducing computational complexity. The proposed system achieves notable improvements in accuracy, precision, and recall, showcasing its potential for real-world emotion recognition applications.

Despite these advancements, certain challenges remain, such as misclassification between similar emotions (neutral vs. sadness) and the limited speaker diversity of the TESS dataset. Future work can explore multimodal approaches (speech + text + facial expressions), transformer-based architectures, and real-world noisy datasets to further enhance performance and generalizability.

In conclusion, the proposed BiLSTM-based SER system, integrated with metaheuristic optimization, provides a robust and scalable solution for emotion detection from speech signals, paving the way for improved human-computer interaction, mental health assessment, and affective computing applications.

FUTURE SCOPE

The future of Speech Emotion Recognition (SER) lies in enhancing its multimodal capabilities, real-world adaptability, and ethical considerations. Integrating text, facial expressions, and physiological signals can improve accuracy, while training on diverse, real-world datasets will make models more robust against background noise and speaker variability. The adoption of pretrained models (e.g., wav2vec, HuBERT) and transfer learning can enable SER systems to work across multiple languages and accents. Further advancements in emotion intensity estimation and context-awareness can enhance applications in mental health monitoring and affective computing.

Additionally, optimizing lightweight deep learning models for edge computing and real-time applications will facilitate deployment on mobile and IoT devices. Future research should also focus on explainable AI (XAI) techniques to ensure transparency, fairness, and ethical AI deployment. Addressing these challenges and opportunities will drive SER towards more accurate, scalable, and socially responsible applications in healthcare, human-computer interaction, and intelligent communication systems.

REFERENCES

[1] Anchit Banga, Bhavik Baheti, Dipesh Sachdev, Yash Jajoo,“Speech Emotion Detection Using State of the Art CNN and LSTM”, International Journal of Innovative Research in Science, Engineering and Technology (IJIRSET) | e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 7.512| Volume 10, Issue 6, June 2021 | DOI:10.15680/IJIRSET.2021.1006329.

[2] Palani Thanaraj Krishnan, Alex Noel Joseph Raj, Vijayarajan Rajangam, “Emotion classification from speech signal based on empirical mode decomposition and non-linear features Speech emotion recognition” Complex and Intelligent Systems (2021) 7:1919-1934 https://doi.org/10.1007/s40747-021-00295-z.

[3] Taiba Majid Wani ,Teddy Surya Gunawan, Syed Asif Ahmad ,Qadri Hasmah Mansor,Mira Kartiwi, Nanang Ismail,“Speech Emotion Recognition using Convolution Neural Networks and Deep Stride Convolutional Neural Networks”, DOI:978-1-7281-7596-6/20/$31.00 ©2020 IEEE.

[4] D.Bharti and P. Kukana, "A Hybrid Machine Learning Model for Emotion Recognition From Speech Signals," 2020 International Conference on Smart Electronics and Communication (ICOSEC), 2020, pp. 491-496, doi: 10.1109/ICOSEC49089.2020.9215376.

[5] H. M. M. Hasan and M. A. Islam, "Emotion Recognition from Bengali Speech using RNN Modulation-based Categorization," 2020 Third International Conference on Smart Systems and Inventive Technology (ICSSIT), 2020, pp. 1131-1136, doi: 10.1109/ICSSIT48917.2020.9214196.

[6] S. Yan, L. Ye, S. Han, T. Han, Y. Li and E. Alasaarela, "Speech Interactive Emotion Recognition System Based on Random Forest," 2020 International Wireless Communications and Mobile Computing (IWCMC), 2020, pp. 1458-1462, doi: 10.1109/IWCMC48107.2020.9148117.

[7] Zhou Qing, college of electrical information, Sichuan University and WangZhong, Wangpeng, College of electrical engineering, and information, Sichuan

University Chengdu university, “Research on speech Emotion recognition technology Based on machine learning”.

[8] Speech Emotion Recognition using Machine Learning R.ANUSHA, P.SUBHASHINI, DARELLI JYOTHI, POTTURI HARSHITHA , JANUMPALLY SUSHMA ,NAMSAMGARI MUKESH Department of Computer Science and Engineering, MLR Institute of Technology, Hyderabad.

[9] Speech Emotion Recognition Based on Speech Segment Using LSTM with Attention Model Bagus Tris Atmaja Japan Adv. Inst. of Science & Tech. and Institut Teknologi Sepuluh Nopember Nomi, Japan and Surabaya, Indonesia bagus@ep.its.ac.id. Masato Akagi School of Information Science Japan Adv. Inst. of Science & Tech. (JAIST) Nomi, Japan akagi@jaist.ac.jp.

[10] Smith, A., & Jones, B. (2023). Ajay Gupta, Siddhesh Moreya, Mukul Sitap, Supriya Chaudhary, “Speech based Emotion Recognition using Machine Learning”, 2021 International Resewrach Journal of Engineering and Technology (IRJET).

[11] https://www.researchgate.net/publication/331442467_A_novel_feature_selection_method_for_speech_emotion_recognition.

[12] Harshit Dolka, Arul Xavier V M, Sujitha Juliet “Speech Emotion Recognition using ANN on MFCC Features”, 2021 3rd International Conference on Signal Processing and Communication (ICPSC), doi:10.1109/ICSPC51351.2021.9451810.

BIOGRAPHIES

Aachal Choudhary is currently pursuing an M.Tech in Artificial Intelligence and Data Science and holds a B.tech in Information Technology from Government College of Engineering, Karad. Her research interests include Machine Learning, Deep Learning and Data Science. Her work explores AI driven solutions with real world applications.