IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Emotion Aware Speech Recognition Using Hybrid Deep Learning With Intelligent Response Generation

[1]Srikanth Gangula , [2]Sabeeha Shaik, [3]Thriveni Vemireddy, [4]Khaja Babu Shaik, [5]Harsha Vardhan Tadigadapa

[1]Associate Professor, [2]Student, [3]Student, [4]Student, [5]Student

[1]Department of Computer Science and Engineering,

[1]Seshadri Rao Gudlavalleru Engineering College, Gudlavalleru, India

Abstract: Recognizing emotions from human speech plays a crucial role in enhancing human–computer interaction and enabling more natural communication systems. This study focuses on the implementation of an explainable and hybrid deep learning-based speech emotion recognition and response generation system, emphasizing the proposed framework combining Convolutional Neural Networks (CNN) and Bidirectional Long Short-Term Memory (BiLSTM) architecture. A selectively compiled dataset was constructed using widely recognized emotional speech datasets, including RAVDESS, EMO-DB, and TESS, followed by the classification of speech samples into multiple emotional categories. The proposed hybrid model utilizes CNN to extract spectral features from Mel spectrograms and BiLSTM to capture temporal emotional variations using MFCC features, further enhanced by the integration of an attention mechanism to highlight emotionally significant segments of speech. The model was trained and optimized using standardized preprocessing techniques and tuning strategies to ensure robustness and reduce bias in performance evaluation. From the experimental analysis, the proposed framework demonstrates improved accuracy compared to individual models while effectively capturing emotional context. Furthermore, to enhance system usability, an emotion-aware response generation module based on the DialoGPT-small language model has been incorporated, enabling the system to produce contextually relevant and empathetic responses. Additionally, a lightweight and user-friendly interface has been developed to allow real-time speech input and visualization of results, thereby facilitating practical deployment in applications such as virtual assistants, customer support systems, and emotion-driven human–computer interaction platforms.

Index Terms - Speech Emotion Recognition, Deep Learning, CNN, BiLSTM, Attention, Emotion-Aware Response, DialoGPT, Human–Computer Interaction

I. INTRODUCTION

Human speech not only conveys information through spoken language but also expresses various emotions such as happiness, sadness, anger, fear, and surprise. These emotional cues play an important role in improving communication and helping individuals understand each other more effectively. When computational systems are capable of recognizing emotions from speech signals, they can interact with humans in a more natural, helpful, and user-friendly manner. Due to this, Speech Emotion Recognition (SER) has become an important area of research in recent years. Earlier approaches to SER primarily relied on hand-crafted audio features combined with traditional machine learning models. Although these methods achieved reasonable performance under controlled conditions, they often failed to generalize

across different speakers, languages, and noisy environments. In addition, many existing systems focus only on emotion classification and do not generate appropriate responses based on the user's emotional state, thereby limiting the effectiveness of human–computer interaction. With the advancement of deep learning techniques, significant improvements have been achieved in speech emotion recognition tasks. Models such as Convolutional Neural Networks (CNNs) have proven effective in extracting meaningful spectral features from Mel spectrogram representations, while Bidirectional Long Short-Term Memory (BiLSTM) networks are capable of capturing temporal emotional variations present in speech signals. Furthermore, attention mechanisms have been introduced to enable the model to focus on emotionally relevant segments of speech while reducing the influence of noise and irrelevant information. Despite these advancements, generating context-aware and emotion-appropriate responses remains a challenging task in intelligent systems. Transformer-based language models such as DialoGPT have demonstrated strong capabilities in producing natural and meaningful conversational responses. Motivated by these developments, this work proposes a hybrid deep learning-based framework for speech emotion recognition and response generation. The proposed system integrates CNN, BiLSTM, and attention mechanisms for accurate emotion classification, followed by an emotion-aware response generation module. The model is trained and evaluated on widely recognized emotional speech datasets, including RAVDESS, EMO-DB, and TESS, to ensure robustness across different speakers and recording conditions.

II. LITERATURE REVIEW

Initially, early work in speech emotion recognition was based on traditional machine learning approaches using acoustic features. Dhole and Kale [1] worked on stress detection in speech using acoustic features and traditional classifiers, where their approach showed moderate accuracy but struggled in noisy real-world conditions. Schuller [2] reviewed the evolution of speech emotion recognition and highlighted the transition from traditional methods to deep learning, along with the importance of datasets and feature extraction. Wahde and Virgolin [3] discussed conversational agents and emphasized the need to incorporate emotion understanding for improved human–computer interaction. With the advancement of deep learning, Mao et al. [4] applied CNN models to extract emotional features from spectrograms, achieving better performance compared to traditional methods, while Indira et al. [5] demonstrated the effectiveness of CNNs in extracting meaningful patterns from complex signal data. Atmaja and Sasou [6] explored bimodal emotion recognition using speech and text, which improved accuracy but increased computational cost. Li et al. [7] applied attention-based techniques to enhance contextual feature learning from speech signals. Zhang et al. [8] introduced a CNN-based model capable of capturing both short-term and long-term temporal patterns in speech data, and Koteswaramma et al. [9] focused on spatiotemporal feature learning to improve classification performance. Zhang et al. [10] further developed a multimodal emotion recognition system achieving high accuracy, though with increased computational requirements. Arora et al. [11] showed that Mel-spectrogram features perform effectively across multiple languages. Keren and Schuller [12] introduced hybrid CNN-RNN models, improving performance by combining spatial and temporal feature extraction. Bahdanau et al. [13] proposed attention mechanisms to enable models to focus on important parts of input data, while Wolf et al. [14] discussed Transformer architectures for efficiently handling long-range dependencies. Li et al. [15] developed emotion-aware chatbots using Transformer-based models. Eyben et al. [16] introduced the GeMAPS feature set for acoustic feature extraction, while Livingstone and Russo [17] created the RAVDESS dataset and Busso et al. [18] introduced the IEMOCAP dataset, both widely used in emotion recognition research. Zhou et al. [19] focused on generating empathetic conversational responses, and Neumann and Vu [20] proposed an attention-based CNN model for improved emotion detection. Fayek et al. [21] concluded that hybrid deep learning models outperform individual models in speech emotion recognition tasks. From the existing studies, it is observed that most systems primarily focus on emotion detection and do not generate appropriate responses based on the user's emotional state. Addressing this limitation, the proposed system utilizes a hybrid CNN–BiLSTM model with an attention mechanism along with a Transformer-based response generator to improve both emotion recognition and response generation.'

III. OBJECTIVES

The primary objective of this work is to design and develop a hybrid deep learning framework that integrates Convolutional Neural Networks (CNN) and Bidirectional Long Short-Term Memory (BiLSTM) architectures, along with multiple attention mechanisms, to enhance the accuracy and reliability of speech emotion recognition. The proposed system aims to automatically extract significant emotional features from speech signals using Mel spectrograms and Mel-Frequency Cepstral Coefficients (MFCCs), thereby eliminating the need for manual feature engineering and reducing the impact of variations caused by different speakers and noisy environments. Furthermore, this work emphasizes the incorporation of attention modules to enable the model to focus on emotionally relevant segments of the speech signal, improving the overall feature representation and classification performance. The system is trained and evaluated using standard emotional speech datasets and assessed through widely used performance metrics, including accuracy, precision, recall, F1-score, and confusion matrix, to ensure comprehensive evaluation across different emotional categories. In addition, this work aims to develop a real-time emotion-aware interaction system capable of detecting the user's emotional state and generating contextually appropriate and empathetic responses using a Transformer-based language model.

IV. METHODOLOGY

The This section describes the overall workflow of the proposed speech emotion recognition system integrated with emotion-aware response generation. The complete workflow consists of several key stages, including dataset collection, audio preprocessing, feature extraction, hybrid model design and training, performance evaluation, and real-time system deployment. Each stage plays a crucial role in ensuring accurate emotion detection and meaningful response generation.

4.1 Proposed System

The proposed system is based on a hybrid deep learning framework that integrates Convolutional Neural Networks (CNN) and Bidirectional Long Short-Term Memory (BiLSTM) networks along with multiple attention mechanisms to accurately identify emotions from speech signals. Once the emotion is recognized, a Transformer-based language model is utilized to generate contextually appropriate and empathetic responses, enhancing the overall interaction experience.

In this framework, the CNN component is responsible for extracting significant spectral features from Mel spectrogram representations, capturing variations in speech characteristics such as pitch and intensity. The BiLSTM component is used to model temporal dependencies and understand how emotional patterns evolve over time using MFCC features. Furthermore, attention layers are incorporated to prioritize emotionally relevant segments of the speech signal while minimizing the influence of noise and irrelevant information. This architecture enables efficient real-time operation, where users can provide speech input through a microphone and receive both emotion predictions and meaningful responses instantly.

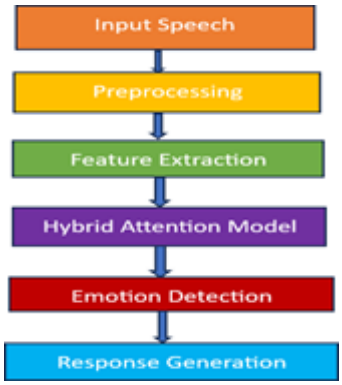

Figure 1: Block Diagram of Proposed System

4.2 System Architecture

The system architecture illustrates the complete processing pipeline from speech input to response generation. It consists of multiple interconnected modules, each responsible for a specific task that contributes to building an intelligent and emotion-aware system:

1. **Speech Input Module**
   This module allows users to provide input either through a microphone or by uploading an audio file.
2. **Preprocessing Unit**
   The input audio is processed to remove noise, normalize amplitude levels, and generate Mel spectrograms and MFCC features.
3. **CNN Feature Extraction Block**
   This module extracts important spectral patterns from speech signals that are associated with emotional expressions.
4. **BiLSTM Feature Extraction Block**
   This component captures temporal variations in speech, enabling the system to understand how emotions change over time.
5. **Attention-Based Feature Fusion (DNN)**
   Features obtained from CNN and BiLSTM are combined and refined using attention mechanisms to emphasize emotionally significant information.
6. **Emotion Classification and Response Generation Layer**
   A Softmax classifier is used to predict the final emotion class, which is then passed to a Transformer-based language model to generate an appropriate emotional response.

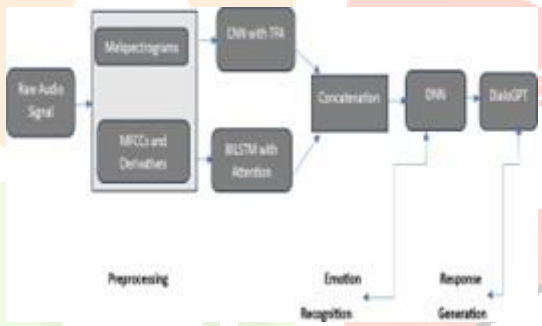

Figure 2: System Architecture

The integration of these modules allows the system to effectively learn both spectral and temporal characteristics of speech signals. This improves performance across different speakers and varying environmental conditions. The final output includes the predicted emotion—such as Happy, Sad, Angry, Fear, Neutral, Disgust, or Surprise—along with a contextually relevant and emotionally appropriate response.

V. IMPLEMENTATION

5.1 Dataset Collection

In this work, emotional speech data is collected by combining widely recognized datasets, including RAVDESS, EMO-DB, and TESS. These datasets contain speech samples representing multiple emotional categories such as neutral, happy, sad, angry, fear, disgust, and surprise. The combined dataset is divided into three subsets, where 80% of the data is used for training, 10% for testing, and 10% for validation. This distribution enables the model to learn effectively while ensuring good generalization on unseen data.

All audio files are organized into separate directories based on emotion categories and dataset splits. Python libraries such as Librosa and PyTorch are used to load and process these audio samples efficiently during training, ensuring consistency and uniformity across the dataset.

5.2 Data Preprocessing

Data preprocessing is performed to standardize all speech samples before they are fed into the model. Initially, all audio recordings are converted to mono format, resampled to 16 kHz, and normalized to minimize variations in loudness. Noise removal and silence trimming techniques are also applied to enhance the clarity of the speech signals.

After preprocessing, important features are extracted from each audio file. Mel spectrograms are used to represent variations in frequency and energy, while MFCC features and their delta coefficients capture temporal vocal patterns associated with emotions. These extracted features are then converted into tensors, enabling efficient processing by the deep learning model during training.

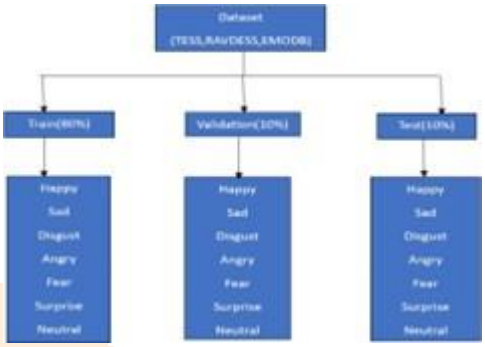

Figure 3: Dataset Organization Structure

5.3 Spectral Feature Learning (CNN Block)

In the first stage of the model, a Convolutional Neural Network (CNN) is used to process Mel spectrogram images. The CNN consists of multiple convolutional layers with ReLU activation functions to learn important spectral patterns such as variations in pitch and energy levels. Max pooling layers are employed to reduce dimensionality while preserving essential information. The extracted features are then flattened and forwarded to the subsequent stages of the model. This enables the CNN to effectively capture emotion-related information from speech frequency patterns.

5.4 Temporal Feature Learning (BiLSTM Block)

The MFCC features are passed into a Bidirectional Long Short-Term Memory (BiLSTM) network to capture temporal dependencies in speech signals. This component enables the model to understand how emotions evolve over time by analyzing variations in speech rate, intensity, and tonal shifts. An attention mechanism is incorporated to highlight emotionally significant segments of the speech signal, thereby improving the overall performance of emotion recognition.

5.5 Hybrid Model Training

The outputs obtained from both the CNN and BiLSTM components are combined in a feature fusion layer. The final emotion is predicted using a Softmax classifier. The model is trained using the Adam optimizer with categorical cross-entropy loss, a batch size of 32, and 40 training epochs. By integrating both spectral and temporal feature learning, the hybrid model reduces misclassification and achieves better performance compared to individual CNN or BiLSTM models.

5.6 Model Evaluation

After training, the performance of the proposed model is evaluated using standard metrics along with a confusion matrix to assess how effectively each emotion category is classified. The training and validation curves are also analyzed to ensure that the model learns appropriately without overfitting. The results indicate that the model achieves strong performance and maintains good generalization when tested on new audio samples that were not part of the training dataset.

5.7 Deployment and Interaction Interface

The final phase of this work involves deploying the trained model as a real-time interactive speech emotion recognition system. The application accepts audio input either through a microphone or via file upload, processes the input to detect the user's emotional state, and generates contextually appropriate empathetic responses using a Transformer-based natural language processing module.

A browser-based graphical user interface is developed to enhance usability, allowing users to interact with the system easily. The interface enables real-time visualization of detected emotions along with the generated responses, providing a seamless and interactive user experience.

Figure 4: Web-Based SER Interface for Audio Input

Figure 5: Real-Time Prediction and Response Display

VI. RESULTS AND DISCUSSION

This section presents the experimental results obtained from the proposed hybrid CNN–BiLSTM model with attention mechanisms for speech emotion recognition. The model was evaluated on a dataset consisting of 6,215 audio samples spanning seven emotion categories: Neutral, Angry, Sad, Fear, Happy, Disgust, and Surprise. Multiple evaluation metrics were used to assess the performance of the system in accurately classifying emotional states from speech signals.

The detailed performance analysis across individual emotion classes demonstrates that the model achieves consistent and reliable results for all seven categories. The use of hybrid feature learning, combining both spectral and temporal information, along with attention mechanisms, contributes to improved classification accuracy and robustness.

Evaluation Metrics

	precision	recall	f1-score	support
neutral	0.99	0.96	0.98	1536
angry	1.00	0.96	0.98	911
sad	0.97	0.97	0.97	846
fear	0.97	0.98	0.98	853
happy	0.95	0.98	0.97	855
disgust	0.96	1.00	0.98	830
surprise	0.98	0.99	0.99	384
accuracy			0.97	6215
macro avg	0.97	0.98	0.98	6215
weighted avg	0.98	0.97	0.97	6215

The training and validation results demonstrate effective learning behavior of the proposed model. The accuracy curves indicate a gradual improvement during the training process, achieving approximately 98% training accuracy and 97% validation accuracy. This close alignment between training and validation performance suggests that the model generalizes well to unseen data. Furthermore, the loss curves show a steady decrease over successive epochs, confirming stable convergence and effective optimization of the model.

Figure 6: Accuracy Curve for Training and Validation

Figure 7: Training and Validation Loss Curves

The training and validation loss curves demonstrate a consistent decrease over epochs, indicating stable convergence of the model during the learning process. The similarity between training and validation loss trends suggests that the model does not suffer from significant overfitting and maintains good generalization capability.

When evaluated on unseen audio samples, the model achieved an overall test accuracy of approximately 97%, indicating strong reliability for real-world applications. The confusion matrix analysis further shows that most emotion classes are correctly identified, with only minimal misclassification between similar emotional states. These results confirm that the proposed hybrid multi-attention model effectively captures both subtle emotional cues and underlying vocal patterns in speech signals.

Overall, the experimental results demonstrate that the proposed system achieves high recognition accuracy and maintains consistent performance across all emotion classes. The model shows strong generalization capability and effectively distinguishes between different emotional states. Additionally, the system provides real-time predictions, making it suitable for interactive voice-based applications.

Figure 8: Confusion Matrix of Proposed Model

VII. CONCLUSION

This work presents a hybrid deep learning–based framework for speech emotion recognition and emotion-aware response generation using speech signals. The proposed system focuses on identifying multiple emotional states from human speech and generating contextually relevant responses, following an end-to-end pipeline comprising audio preprocessing, feature extraction using a CNN–BiLSTM architecture, attention-based feature enhancement, and real-time deployment. The hybrid model effectively combines Convolutional Neural Networks (CNN) for extracting spectral features and Bidirectional Long Short-Term Memory (BiLSTM) networks for capturing temporal emotional variations, enabling the system to learn both local and sequential characteristics of speech signals. The integration of attention mechanisms further enhances the model's ability to focus on emotionally significant segments, improving overall classification performance. Experimental results demonstrate that the proposed framework achieves high accuracy of approximately 97%, along with strong evaluation performance

across multiple emotion classes, indicating its robustness under varying speech conditions. The confusion matrix analysis further validates the model's capability to reliably distinguish between different emotional states. In addition to emotion classification, the integration of a Transformer-based language model enables the generation of contextually appropriate and empathetic responses, thereby enhancing the naturalness of human–computer interaction. Furthermore, the system has been deployed using a real-time interactive interface, allowing users to input speech and receive unified emotion predictions along with meaningful conversational responses. The proposed framework demonstrates effectiveness, robustness, and practical applicability in developing emotionally intelligent systems, including virtual assistants, healthcare support systems, learning assistants, and empathetic AI companions. Future work will focus on improving model generalization across diverse languages and accents, expanding emotion categories, and optimizing response generation for more personalized and context-aware interactions.

REFERENCES

[1] N. Dhole and S. Kale, "Stress detection in speech using machine learning and AI techniques," in Proc. ICMLIP, Springer, pp. 11–26, 2020.

[2] B. W. Schuller, "Speech emotion recognition: Two decades in a nutshell, benchmarks, and ongoing trends," Commun. ACM, vol. 61, no. 5, pp. 90–99, 2018.

[3] M. Wahde and M. Virgolin, "Conversational agents and their applications," in Handbook on Computer Learning and Intelligence, vol. 2, World Scientific, pp. 497–544, 2022.

[4] Q. Mao, M. Dong, Z. Huang, and Y. Zhan, "Learning features for speech emotion recognition using CNN," IEEE Trans. Multimedia, vol. 16, no. 8, pp. 2203–2213, 2014.

[5] D. N. V. S. L. S. Indira et al., "Detection of cardiac arrhythmia using multi-perspective convolutional neural network for ECG heartbeat classification," Revue d'Intelligence Artificielle, vol. 36, no. 4, pp. 629–634, Aug. 2022.

[6] B. T. Atmaja and A. Sasou, "A survey on bimodal speech emotion recognition using audio and language features," Speech Commun., vol. 140, pp. 11–28, 2022.

[7] R. Li, Z. Wu, J. Jia, S. Zhao, and H. Meng, "Dilated residual networks with multi-head attention for emotion recognition," in Proc. IEEE ICASSP, pp. 6675–6679, 2019.

[8] S. Zhang, S. Zhang, T. Huang, and W. Gao, "Speech emotion recognition with deep CNN and temporal pyramid matching," IEEE Trans. Multimedia, vol. 20, no. 6, pp. 1576–1590, 2017.

[9] A. Koteswaramma, M. Babu Rao, and G. Jaya Suma, "Adaptive learning framework using spatiotemporal features," Signal Image Video Process., Jan. 2024.

[10] S. Zhang et al., "Multimodal emotion recognition combining audio, video, and text data," Expert Syst. Appl., 2023.

[11] S. Arora et al., "Multilingual speech emotion recognition using Mel-spectrogram features," in Proc. IEEE ICASSP, pp. 5260–5264, 2021.

[12] T. Keren and B. W. Schuller, "CNN-RNN for speech emotion recognition," in Proc. IEEE ICASSP, pp. 112–116, 2016.

[13] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation with attention mechanism," in Proc. ICLR, 2015.

[14] T. Wolf et al., "Transformers: State-of-the-art natural language processing," in Proc. EMNLP, pp. 38–45, 2020.

[15] D. Li, X. Wu, and S. Zhang, "Emotion-aware chatbot using Transformer models," IEEE Access, vol. 10, pp. 46754–46767, 2022.

[16] F. Eyben et al., "GeMAPS: Standard acoustic features for emotion recognition," IEEE Trans. Affective Comput., vol. 7, pp. 190–202, 2016.

[17] S. R. Livingstone and F. A. Russo, "The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS)," PLOS ONE, vol. 13, no. 5, 2018.

[18] C. Busso et al., "IEMOCAP: Interactive emotional motion capture database," Lang. Resour. Eval., vol. 42, pp. 335–359, 2008.

[19] H. Zhou et al., "Empathy-driven emotional conversation generation," in Proc. ACL, pp. 1171–1182, 2020.

[20] D. Neumann and N. T. Vu, “Attention-based CNN for speech emotion recognition,” in Proc. Interspeech, pp. 1228–1232, 2017.

[21] A. Vaswani et al., “Attention is all you need,” in Proc. NeurIPS, pp. 5998–6008, 2017.

[22] H. Fayek, M. Lech, and L. Cavedon, “Evaluating deep learning architectures for speech emotion recognition,” Neural Netw., vol. 92, pp. 60–68, 2017.