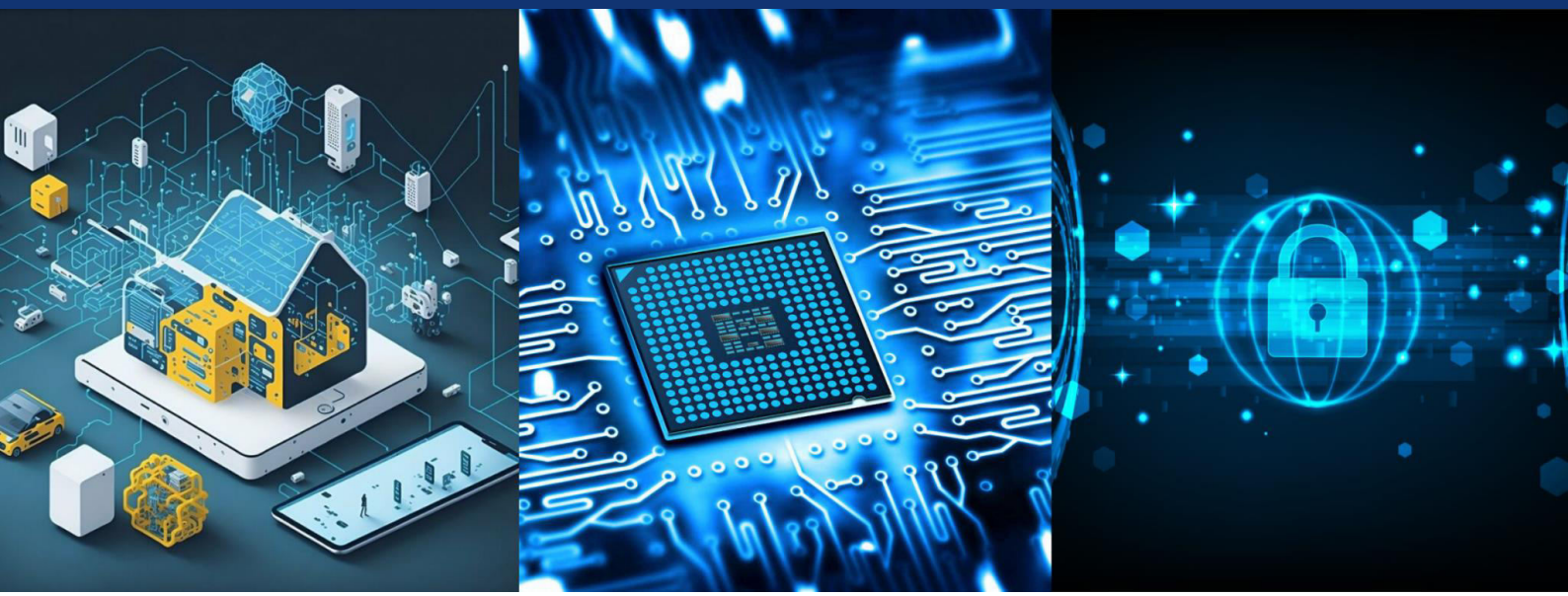
IJIRCCE

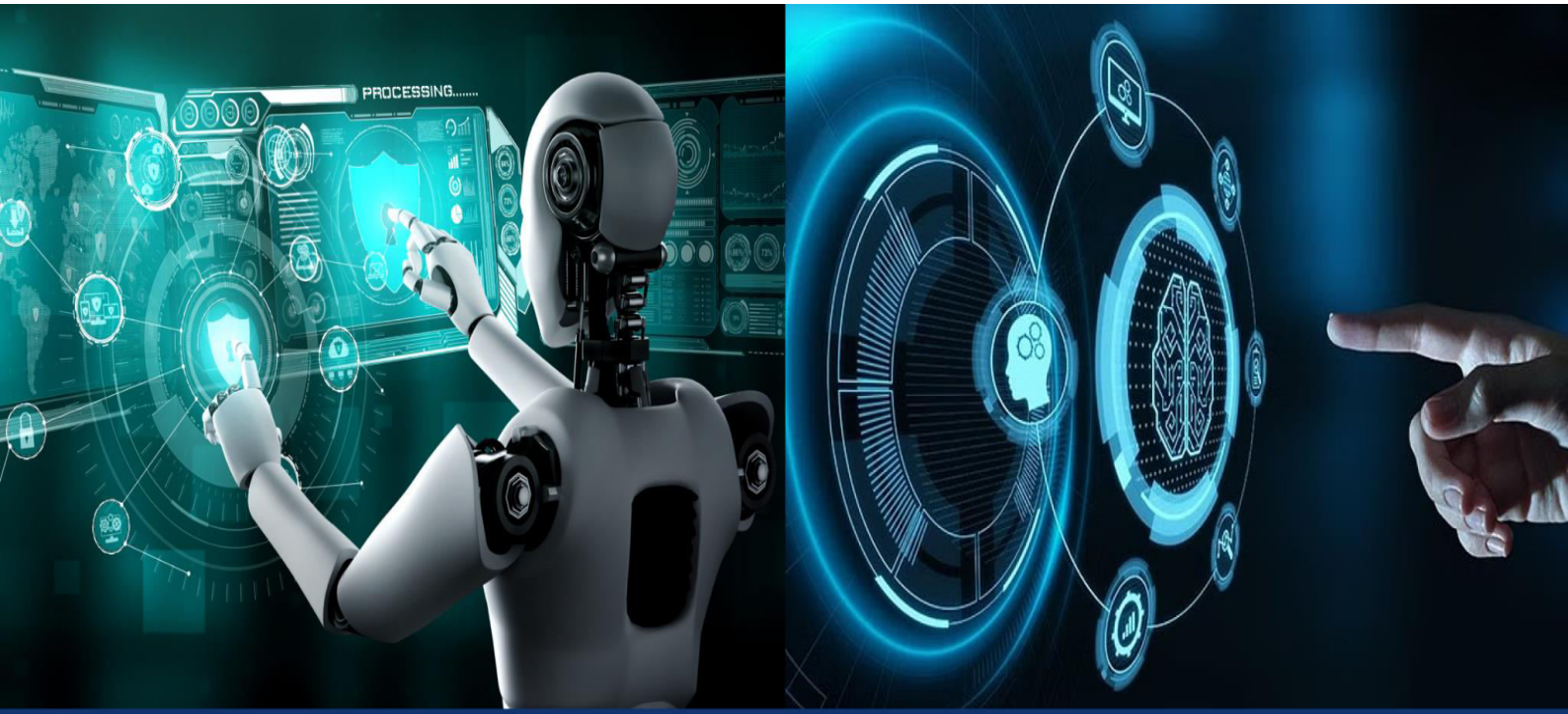
ISSN(O): 2320-9801

ISSN(P): 2320-9798



International Journal of Innovative Research in Computer and Communication Engineering

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



Impact Factor: 9.927

Volume 14, Issue 6, June 2026

© 2026 IJIRCCE | Volume 14, Issue 6, June 2026 DOI: 10.15680/IJIRCCE.2026.1406073

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 9.927| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

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

Speech Emotion Recognition in Human Voice

Ms. K. Vaishnavi[1], Mrs. M. Vijayalakshmi[2]

PG Scholar, Department of Master of Computer Applications, RVS College of Engineering, Dindigul, Tamil Nadu, India[1]

Assistant Professor, Department of Master of Computer Applications, RVS College of Engineering, Dindigul, Tamil Nadu, India[2]

ABSTRACT: Speech Emotion Recognition (SER) is a rapidly evolving area in the field of human-computer interaction and affective computing. It focuses on the automatic identification of emotions embedded in human speech, enabling machines to interpret not just what is said, but how it is said. This capability is vital in developing emotionally intelligent systems, with applications ranging from virtual assistants and customer service bots to healthcare diagnostics and security systems. This paper presents the design and implementation of a SER system that leverages deep learning and robust signal processing to recognize emotional states from speech in real time. Key acoustic features including MelFrequency Cepstral Coefficients (MFCCs), pitch, energy, and zero-crossing rate are extracted and fed into deep neural networks for emotion classification. The system is evaluated on publicly available emotional speech databases and demonstrates improved accuracy over traditional machine-learning-based approaches.

I. INTRODUCTION

Speech is one of the most natural and expressive forms of human communication. Beyond conveying words and sentences, speech carries a rich spectrum of emotional information happiness, sadness, anger, fear, disgust, surprise, and neutrality embedded in acoustic patterns such as pitch, rhythm, tempo, and intensity. The automatic recognition of these emotions from speech signals, known as Speech Emotion Recognition (SER), has gained significant research attention over the past two decades.

SER plays a pivotal role in developing emotionally aware machines. Emotionally intelligent systems canimprove human-computer interaction by responding appropriately to a user's emotional state, enhance customer satisfaction in call centers by detecting frustration early, assist in mental health monitoring by flagging stress or depression indicators, and support elderly care systems by detecting distress. As speech becomes the primary interface for interacting with AI assistants, smart devices, and cloud services, the need for accurate and robust SER systems becomes increasingly urgent.

II. EXISTING SYSTEM

.Traditional Speech Emotion Recognition systems primarily rely on handcrafted acoustic features and classical machine learning techniques to identify emotional states from speech signals. One major limitation of existing systems is their inability to effectively model the temporal dynamics of emotions in speech. Emotions evolve over time, but conventional approaches of ten treat speech segments as independent frames, ignoring long-term dependencies. Traditional systems are highly sensitive to noise , variations in speaking style, accent, and recording conditions. Most existing datasets are limited in size and recorded under ideal laboratory conditions, leading to poor performance when deployed in spontaneous or noisy environments.

Disadvantages of Existing Systems

- Low recognition accuracy on noisy, blurred, or real-world audio environments
- Inability to effectively model the temporal dynamics of emotions in speech
- Highly sensitive to noise, variations in speaking style, accent, and recording conditions
- Most existing datasets are limited in size and recorded under ideal laboratory conditions

© 2026 IJIRCCE | Volume 14, Issue 6, June 2026 DOI: 10.15680/IJIRCCE.2026.1406073

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 9.927| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

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

III. PROPOSED SYSTEM

The proposed Speech Emotion Recognition system employs a multi-stage pipeline that begins with raw audio acquisition and ends with emotion category prediction. The system is designed for real-time operation and integrates with web-based applications through a Flask backend.

Advantages of the Proposed System

- Significantly improved emotion recognition accuracy through deep learning
- Effective modeling of temporal dynamics using Bidirectional LSTM layers
- Attention mechanism focuses on emotionally salient speech segments.• Real-time emotion detection integrated via Flask web backend

IV. SYSTEM ARCHITECTURE

The system architecture of the proposed Speech Emotion Recognition system follows a fourstage modular pipeline. The pipeline begins with the Input Dataset and Training Dataset, which are passed through the Pre-Processing stage to clean and normalize the audio signals. The preprocessed data is then sent to the Feature Extraction module, where key acoustic features such as MFCCs, pitch, and energy are extracted. These features are fed into the Hybrid Model (CNN-BiLSTM), which classifies the speech into one of seven emotion categories: Angry, Happy, Disgusting, Sad, Neutral, Fear, and PS. In parallel, live Speech Signal input from a microphone is also preprocessed and passed through feature extraction before being classified by the same hybrid model in real time.

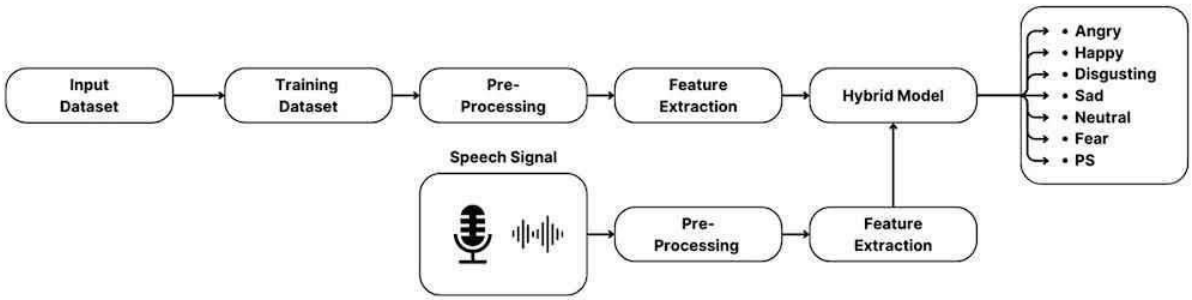


Figure 1: System Architecture – Speech Emotion Recognition Pipeline

V. MODULES DESCRIPTION

The system is composed of four principal modules, each forming a distinct stage of the emotion recognition pipeline. The modular architecture promotes maintainability, scalability, and ease of integration with external systems.

1.1 Data Collection Module

The Data Collection Module is the entry point of the SER pipeline. It is responsible for gathering audio data from two primary sources:

1. live streaming from microphones in call center environments, and
2. prerecorded audio files from existing call logs or public emotional speech databases.

For live data, the module captures audio at a sampling rate of 22,050 Hz in WAV format to ensure sufficient signal fidelity for downstream feature extraction. It concurrently records relevant metadata including call duration, timestamp, agent identifier, and customer session ID, storing all data in a structured database for audit and compliance purposes.

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 9.927| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

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

The module is designed for scalability and integrates with cloud storage services such as AWS S3 or Google Cloud Storage to handle large volumes of concurrent audio streams.

1.2 Preprocessing Module

Raw audio signals collected from real-world environments are often contaminated by background noise, microphone artifacts, electrical interference, and channel distortions. The Preprocessing Module addresses these issues through a multi-step signal cleaning and normalization pipeline.

The preprocessing pipeline consists of the following stages:

1. Noise Reduction: Spectral subtraction and Wiener filter-based noise reduction algorithms are applied to attenuate background interference while preserving speech clarity.
2. Voice Activity Detection (VAD): Silence and non-speech segments are identified and removed using energy-threshold and zero-crossing-rate-based VAD, reducing computational overhead in subsequent stages.

1.3 Feature Extraction Module

The Feature Extraction Module transforms preprocessed audio frames into compact numerical representations that capture the emotional characteristics of speech. Feature selection is guided by psychoacoustic and linguistic research on the acoustic correlates of emotion.

The following feature types are extracted:

- Mel-Frequency Cepstral Coefficients (MFCCs): 40 MFCC coefficients along with their first-order (delta) and second-order (delta-delta) derivatives are extracted per frame. MFCCs are the most widely used features in SER as they effectively represent the spectral envelope of speech, which carries significant emotional information.
- Pitch (Fundamental Frequency, F0): The pitch contour is extracted using the autocorrelation method. Pitch variation is a strong indicator of emotional arousal high arousal emotions such as anger and excitement are characterized by elevated and variable pitch, while low arousal states such as sadness show flatter, lower pitch profiles.

1.4 Emotion Classification Module

The Emotion Classification Module takes the extracted feature vectors as input and outputs a probability distribution over the seven target emotion categories: happiness, sadness, anger, fear, disgust, surprise, and neutrality. The predicted emotion is the category with the highest posterior probability.

The classification model is a hybrid CNN-BiLSTM architecture:

- Convolutional Layers: Three 1D convolutional layers with ReLU activations and maxpooling extract local temporal patterns and short-range feature correlations from the MFCC sequence.
- Bidirectional LSTM Layers: Two stacked BiLSTM layers with 128 hidden units each capture long-range temporal dependencies and forward/backward contextual information in the speech sequence.
- Attention Mechanism: A self-attention layer assigns importance weights to different time steps, allowing the model to focus on emotionally expressive segments of the utterance.

VI. RESULTS AND DISCUSSION

The proposed SER system was evaluated on the RAVDESS and TESS datasets using a stratified 80/20 train-test split with 5-fold cross-validation to ensure reliable performance estimation. The evaluation metrics include overall classification accuracy, weighted F1-score, and per-emotion precision and recall.

© 2026 IJIRCCE | Volume 14, Issue 6, June 2026 DOI: 10.15680/IJIRCCE.2026.1406073

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 9.927| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

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

1.5 Performance Comparison

Table 1: Classification Accuracy Comparison Across Models (RAVDESS Dataset)

Model	Accuracy (%)	F1-Score	Notes
SVM (Baseline)	68.4	0.671	Handcrafted features only
Random Forest	71.2	0.704	Ensemble, no temporal modeling
Standard CNN	74.8	0.739	Local feature learning
LSTM	76.3	0.756	Temporal modeling only
CNN-BiLSTM (Proposed)	83.6	0.829	Hybrid + Attention

VII. SCREENSHOTS

Screenshot – Speech Emotion Recognition Use cases

Figure 2 presents a screenshot illustrating the real-world use cases of the Speech Emotion Recognition system for cybersecurity applications. The screenshot shows two key application areas enhancing biometrics to prevent identity theft and detecting fraud to prevent security incidents.

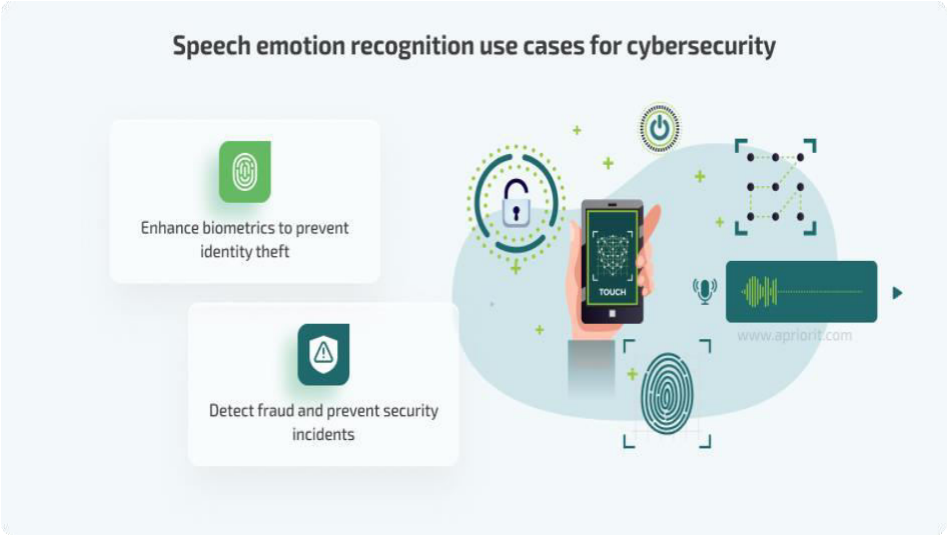


Figure 2: Screennshot – Speech Emotion Recognition Use Cases for Cybersecurity

© 2026 IJIRCCE | Volume 14, Issue 6, June 2026 DOI: 10.15680/IJIRCCE.2026.1406073

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 9.927| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

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

Screenshot – Waveplots, Spectrograms and Emtion Classification

Figure 3 presents a screenshot of the system output displaying the audio waveplots and corresponding spectrograms for four emotion categories. The waveplot column shows the raw audio signal patterns, the spectrogram column displays the frequency-time representation, and the emotions column shows the predicted labels Happy, Sad, Angry, and Fearful confirming successful emotion classification by the CNN-BiLSTM model.

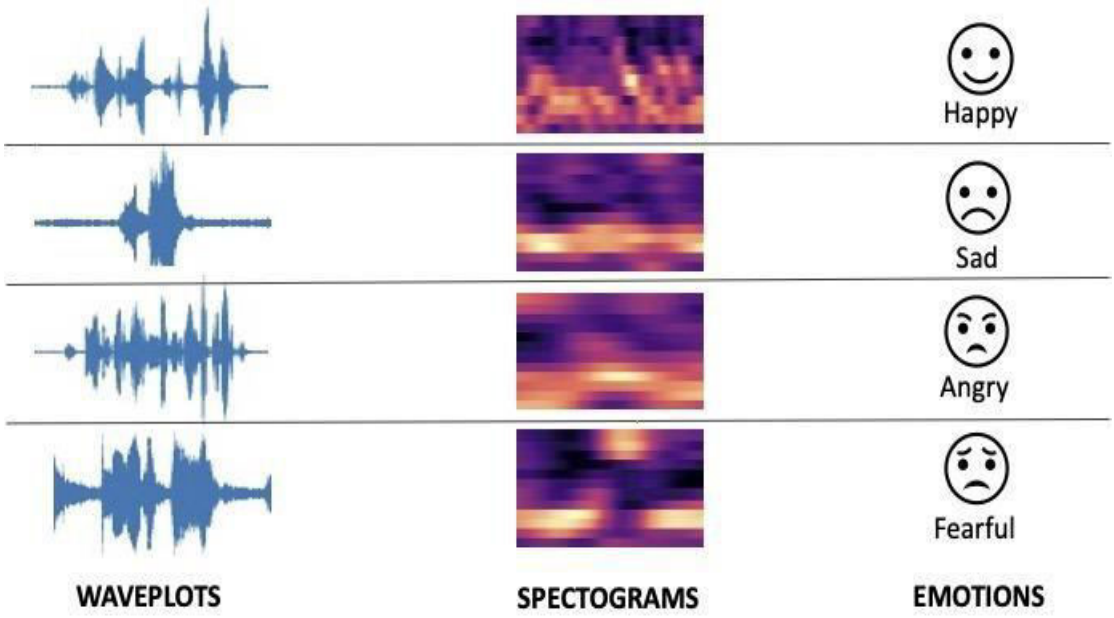


Figure 3: Screenshot – Waveplots, Spectrograms and Emotion Classification Output

Screenshot – MFCC Spectrogram Visualization of Speech Emotions

Figure 4 displays four MFCC-based spectrogram plots of emotional speech samples. The horizontal axis represents time, the vertical axis represents frequency (0–8192 Hz), and color intensity indicates energy level. Dark blue regions indicate low energy, while yellow and orange indicate high energy. Each spectrogram reflects a distinct emotional pattern higharousal emotions such as anger and happiness show dense energy activity, whereas low-arousal emotions such as sadness and neutrality display sparse patterns. These visualizations validate the ability of the proposed CNN-BiLSTM model to distinguish emotional states through spectral features.

© 2026 IJIRCCE | Volume 14, Issue 6, June 2026 DOI: 10.15680/IJIRCCE.2026.1406073

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 9.927| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

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

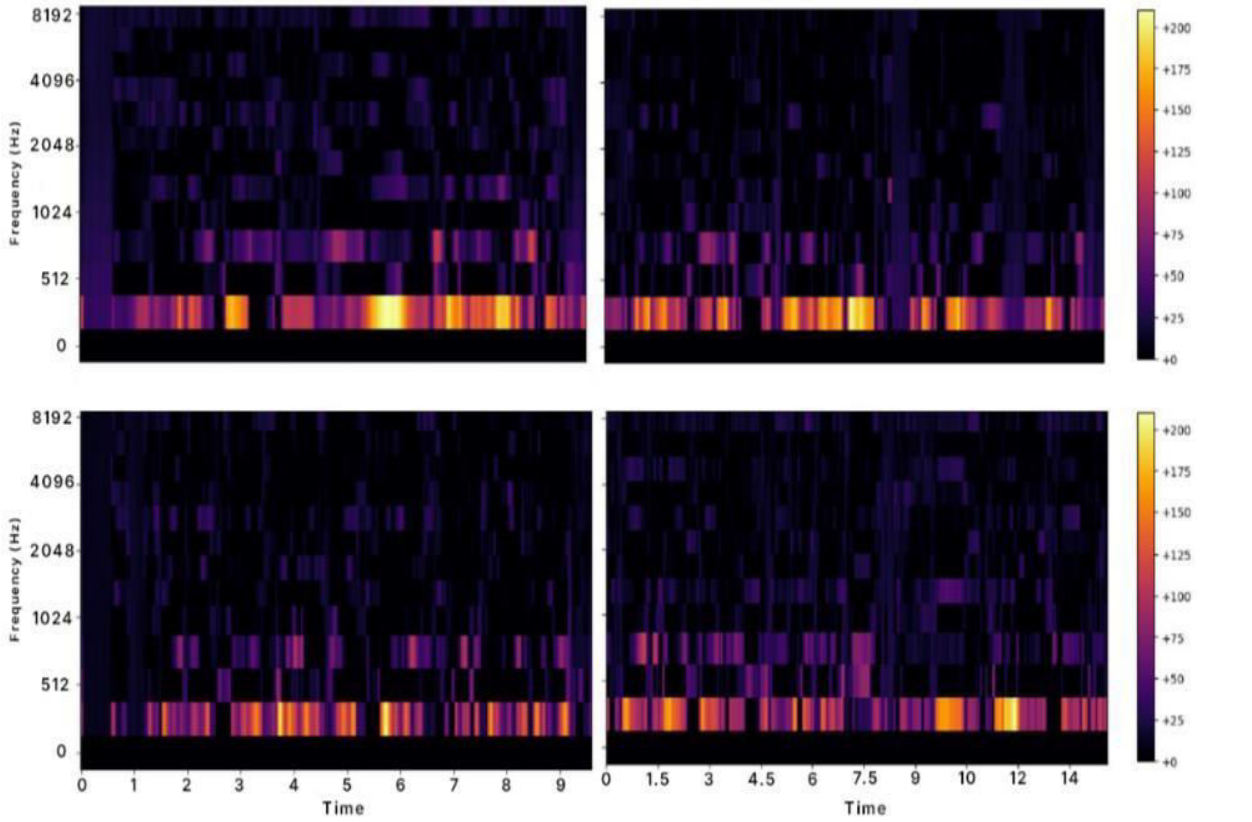


Figure 4: Screenshot – MFCC Spectrogram Visualization

VIII. CONCLUSION

This paper has presented a comprehensive Speech Emotion Recognition system designed for real-world deployment in call center environments and other human-computer interaction contexts. The proposed system addresses the key limitations of traditional SER approaches through a robust preprocessing pipeline, rich multidimensional feature extraction, and a hybrid CNN-BiLSTM deep learning architecture augmented with attention mechanisms. The implementation of the proposed AI-based Speech Emotion Analysis system enhances service quality, employee performance, and customer satisfaction by analyzing voice interactions in real time

IX. FUTURE ENHANCEMENTS

While the proposed CNN-BiLSTM system demonstrates strong performance on benchmark datasets, several promising directions remain for future research and system improvement. One key area of future work is the development of multilingual and cross-cultural emotion recognition capabilities. Current models are predominantly trained on English-language datasets, limiting their generalizability across different languages and cultural expressions of emotion. Future iterations of the system will incorporate multilingual corpora and transfer learning strategies to bridge this gap.

REFERENCES

1. S. G. Koolagudi and K. S. Rao, "Emotion recognition from speech: A review," International Journal of Speech Technology, vol. 15, no. 2, pp. 99-117, 2012.
2. R. W. Picard, Affective Computing. MIT Press, 1997.
3. B. Schuller, "Speech emotion recognition: Two decades in a nutshell, benchmarks, and ongoing trends," Communications of the ACM, vol. 61, no. 5, pp. 90-99, 2018.

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 9.927| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

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

4. M. El Ayadi, M. S. Kamel, and F. Karray, "Survey on speech emotion recognition: Features, classification schemes, and databases," Pattern Recognition, vol. 44, no. 3, pp. 572-587, 2011.[5] S. S. Kanojia, "Speech emotion recognition for call center conversations," International Journal of Engineering Research & Technology, vol. 9, no. 7, pp. 45-50, 2020.
5. D. Low, K. Bentley, and S. Ghosh, "Automated assessment of psychiatric disorders using speech: A systematic review," Journal of Medical Internet Research, vol. 22, no. 5, p. e17101, 2020.
6. R. N. Naoum, "Real-time emotion recognition for security systems," IEEE Security & Privacy, vol. 18, no. 4, pp. 34-42, 2020.
7. M. H. Zafar, "Driver emotion recognition for intelligent vehicles," IEEE Transactions on Intelligent Transportation Systems, vol. 22, no. 8, pp. 4824-4835, 2021.
8. Z. Huang, M. Dong, and Q. Mao, "Speech emotion recognition using CNN with attention mechanism," IEEE Access, vol. 7, pp. 145421-145432, 2019.
9. R. A. Khalil, E. Jones, M. I. Babar, T. Jan, M. H. Zafar, and T. Alhussain, "Speech emotion recognition using deep learning techniques: A review," IEEE Access, vol. 7, pp. 117327117345, 2019.

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INNO SPACE SJIF Scientific Journal Impact Factor

doi crossref

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

9940 572 462 6381 907 438 ijircce@gmail.com

www.ijircce.com

Scan to save the contact details