

Real-Time Speech Emotion Recognition Using Deep Learning Techniques

TAMIZHARASAN K, M.E. Students

Department of Computer Science and Engineering, MRK Institution, Kattumannarkoil, India

Mrs.D.SUGUNA, Assistant Professor

Department of Computer Science and Engineering, MRK Institution of Technology, Kattumannarkoil, India

Mrs.T.KALAIARASI, HOD

Department of Computer Science and Engineering, MRK Institution of Technology, Kattumannarkoil, India

Abstract--Speech Emotion Recognition (SER) is an important area of affective computing that enables machines to identify human emotional states from speech signals. In natural communication, emotions are expressed through pitch, tone, rhythm, intensity, speech rate, and spectral variations. However, recognizing these emotional cues automatically is difficult because speech is affected by speaker identity, accent, recording condition, background noise, and temporal changes. This paper presents an IEEE-style expanded article on a real-time SER system using deep learning techniques. The proposed system uses benchmark emotional speech datasets including TESS, EmoDB, and RAVDESS. Acoustic and spectral features such as Mel Frequency Cepstral Coefficients (MFCC), Zero Crossing Rate (ZCR), Mel spectrogram, Root Mean Square (RMS) energy, and Chroma features are extracted from speech signals. To improve generalization, the system applies augmentation methods including noise addition, pitch shifting, time stretching, and spectrogram shifting. Several models are evaluated, including Multilayer Perceptron (MLP), Convolutional Neural Network (CNN), ResNet50, and a hybrid CNN+Bidirectional Long Short-Term Memory (Bi-LSTM) model. Experimental results show that the proposed CNN+Bi-LSTM model achieves the highest performance with 95.42% accuracy, 94.90% precision, 94.62% recall, 94.75% F1-score, and 620 ms inference latency. The results confirm that combining CNN-based spatial spectral learning with Bi-LSTM-based temporal sequence learning improves real-time speech emotion recognition under practical conditions.

Index Terms--Speech emotion recognition, deep learning, CNN, Bi-LSTM, MFCC, Mel spectrogram, data augmentation, affective computing, real-time emotion recognition.

I. INTRODUCTION

Human communication is not limited to the literal meaning of words. A large part of meaning is carried by vocal expression, tone, intensity, rhythm, pitch variation, and speaking rate. A person may speak the same sentence with happiness, anger, fear, sadness, or neutrality, and listeners can usually recognize the emotional state from vocal cues. Machines, however, do not naturally understand these cues. This limitation has motivated the development of Speech Emotion Recognition (SER), which attempts to classify human emotions from speech signals using computational methods.

SER is a major component of affective computing and human-computer interaction. Affective computing aims to build systems that can recognize, interpret, and respond to human emotions. With the rapid growth of artificial intelligence, virtual assistants, online education, telemedicine, customer support automation, driver safety systems, and social robotics, emotion-aware systems are becoming increasingly necessary. A system that recognizes whether a user is angry, sad,

frustrated, calm, or happy can respond in a more natural and supportive manner.

Speech is particularly important for emotion recognition because it can be captured easily through a microphone without requiring a camera or additional wearable sensors. It also provides continuous information about the speaker during natural interaction. Unlike text, speech preserves prosody and vocal dynamics. These characteristics make SER useful for real-time applications such as call center monitoring, mental health screening, smart learning platforms, and intelligent personal assistants.

Despite its importance, SER remains a challenging task. Emotional speech varies greatly across speakers, languages, genders, accents, age groups, and recording environments. Background noise, microphone quality, room reverberation, and emotional overlap can degrade recognition performance. For example, fear and surprise may share high pitch and fast speech patterns, while sadness and neutrality may both contain low energy and slow speaking patterns. An accurate SER system must therefore learn both spectral information and temporal emotional transitions.

In earlier phases of SER research, classical machine learning models used handcrafted features such as MFCC, energy, pitch, and ZCR with classifiers such as Support Vector Machine, K-Nearest Neighbor, Hidden Markov Model, and Gaussian Mixture Model. Although these methods are useful, they depend heavily on manual feature design and often fail to generalize under real-world acoustic variability. Deep learning methods overcome many of these limitations by learning hierarchical representations directly from speech features.

Phase I of the present work implemented baseline deep learning models including MLP, CNN, and ResNet50. These models achieved promising results, but several limitations were identified. CNN and ResNet50 extracted spatial features from spectrogram representations but did not fully capture temporal sequence relationships in emotional speech. In addition, deeper architectures such as ResNet50 required more computational resources and showed higher inference latency during real-time use. To address these limitations, Phase II proposes a hybrid CNN+Bi-LSTM model that learns both spectral-spatial and temporal emotional features.

The contribution of this article is an expanded IEEE-style presentation of the Phase II SER framework. The article reorganizes the thesis content into a research-paper structure, explains the system pipeline, describes datasets and features, presents the hybrid architecture, includes tables and charts, and discusses the experimental results. The goal is to provide a clear, professional, and publication-ready article format suitable for academic submission or presentation.

II. MOTIVATION AND PROBLEM STATEMENT

The motivation for this study arises from the increasing need for emotionally intelligent systems. Many digital platforms can recognize speech content, but they cannot reliably understand the emotional state behind the speech. A virtual assistant may answer a question correctly but still fail to respond appropriately to frustration. A call center system may transcribe a customer complaint but fail to detect anger or dissatisfaction. An online learning platform may deliver lessons but fail to recognize boredom or confusion. SER can reduce this gap by making systems more sensitive to human feelings.

Emotion recognition from speech can also support health and safety applications. In healthcare, abnormal emotional patterns may help monitor stress, anxiety, or depressive tendencies. In driver monitoring, anger,

fatigue, and stress detection may reduce accident risk. In education, emotional feedback can help teachers and adaptive learning systems understand learner engagement. In customer service, detecting dissatisfaction can help supervisors intervene at the right time. These examples demonstrate why real-time SER is practically important.

The main problem is that real-world speech emotion recognition must operate under difficult conditions. Speech signals may contain background noise, silence, speaker variability, and recording inconsistency. Traditional models often perform well on clean laboratory datasets but fail when exposed to real microphone input. Deep CNN models can learn spatial patterns from spectrograms, but they may ignore temporal dependencies. Emotional information unfolds over time, so a single static representation may not capture the complete emotional context.

Therefore, the core problem addressed in this paper is the development of a robust, accurate, and efficient real-time SER system that can learn both spectral and temporal characteristics from speech. The system must also handle noise and speaker variation, reduce overfitting, maintain low inference latency, and generalize across multiple benchmark datasets. The proposed CNN+Bi-LSTM architecture with data augmentation is designed to solve this problem.

III. RELATED WORK

Early SER systems were based on handcrafted acoustic descriptors. Features such as MFCC, pitch, energy, formants, ZCR, and voice quality parameters were extracted from speech and then classified using statistical or classical machine learning models. These approaches provided a foundation for SER research but required careful manual feature engineering. Their performance depended strongly on the quality of feature selection and they often struggled with cross-speaker and cross-dataset generalization.

Deep learning changed the direction of SER research by allowing models to learn abstract representations from speech data. CNN-based models became popular because spectrograms can be treated like two-dimensional images. A Mel spectrogram contains frequency and time information, and CNN filters can capture local patterns such as changes in energy, harmonic structure, and spectral texture. CNN models therefore improve feature learning compared with handcrafted approaches.

However, speech is sequential by nature. Emotional patterns are not only visible in isolated frames but also

in transitions across time. Recurrent Neural Networks (RNN), Long Short-Term Memory (LSTM), Gated Recurrent Unit (GRU), and Bi-LSTM architectures are used to model temporal dependencies. Bi-LSTM is particularly useful because it processes sequences in both forward and backward directions, allowing the model to learn context from past and future frames.

Hybrid CNN-RNN models have become common in recent SER research because they combine the strengths of both architectures. CNN layers extract local spectral patterns, while recurrent layers model temporal progression. Research surveys and experimental studies have shown that hybrid models can outperform standalone CNN or traditional classifiers, especially when the task involves dynamic emotional speech.

Data augmentation has also become an important strategy in SER because emotional speech datasets are usually limited in size. Techniques such as noise injection, pitch shifting, time stretching, and spectrogram transformation expose the model to diverse speech variations. This reduces overfitting and improves robustness when the system is deployed in real acoustic environments.

The reviewed literature indicates that an effective real-time SER system should use a strong feature representation, a hybrid deep learning model, augmentation for robustness, and evaluation metrics beyond simple accuracy. The proposed work follows these principles by combining MFCC, Mel spectrogram, ZCR, RMS, and Chroma features with a CNN+Bi-LSTM architecture and augmentation-based training.

IV. PROPOSED METHODOLOGY

The proposed SER methodology follows a structured deep learning pipeline. The major steps are speech acquisition, preprocessing, feature extraction, data augmentation, model training, model evaluation, and real-time inference. The pipeline accepts either dataset-based speech recordings or live microphone input. It then standardizes the speech signal, converts it into useful feature representations, trains deep learning models, and produces an emotion label with confidence information.

Fig. 1 shows the overall system architecture extracted from the thesis implementation. The architecture begins with speech input from a dataset or live microphone, followed by audio preprocessing, feature extraction, data augmentation, CNN layers, Bi-LSTM layers, dense classification layers, and real-time emotion output. This

complete pipeline is designed for both offline model training and real-time prediction.

The methodology emphasizes the combination of spatial and temporal learning. The CNN component receives spectrogram-based representations and learns local frequency-time patterns. The Bi-LSTM component then interprets the sequence of extracted features to understand emotional progression. Finally, dense and Softmax layers classify the speech sample into one of the predefined emotional categories.

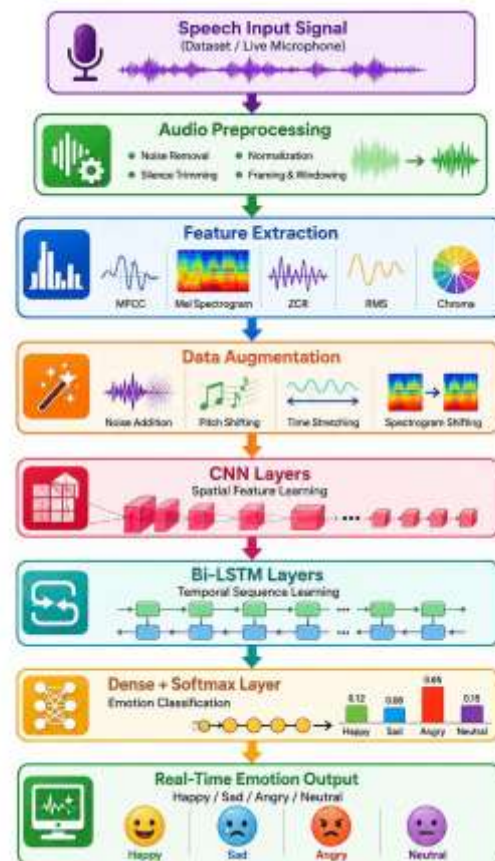


Fig. 1. Overall architecture diagram of the proposed real-time SER system.

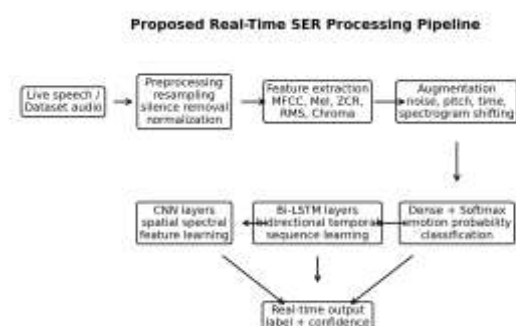


Fig. 2. Simplified processing workflow used in the proposed SER pipeline.

V. DATASET DESCRIPTION

The system was trained and evaluated using three benchmark emotional speech datasets: TESS, EmoDB, and RAVDESS. The use of multiple datasets improves

diversity because each dataset contains different speakers, recording styles, emotional categories, and acoustic conditions. Combining these datasets helps the model learn more general emotional patterns instead of becoming dependent on a single source.

The Toronto Emotional Speech Set (TESS) contains approximately 2800 utterances recorded by two female speakers. It includes emotional categories such as anger, disgust, fear, happiness, pleasant surprise, sadness, and neutral speech. TESS is useful because the recordings are clean and well-balanced across emotions, making it suitable for supervised learning.

The Berlin Emotional Database (EmoDB) contains 535 emotional utterances recorded by ten German speakers. It includes both male and female voices and contains emotions such as anger, boredom, disgust, fear, happiness, sadness, and neutrality. EmoDB contributes language and speaker diversity, which helps evaluate whether the model can handle emotional expression beyond a single dataset.

RAVDESS is a large audio-visual emotional speech and song dataset recorded by professional actors. The thesis document describes it as containing more than 7000 emotional recordings and including emotional states such as calm, happy, sad, angry, fearful, disgust, surprised, and neutral. Its large size, multiple speakers, gender diversity, and emotional intensity variations make it valuable for training robust deep learning models.

TABLE I
SUMMARY OF DATASETS USED IN THE STUDY

Dataset	Approx. size	Speakers	Emotions / characteristics
TESS	2800 utterances	2 female speakers	Anger, disgust, fear, happiness, surprise, sadness, neutral
EmoDB	535 utterances	10 German speakers	Anger, boredom, disgust, fear, happiness, sadness, neutral
RAVDESS	More than 7000 recordings	24 professional actors	Calm, happy, sad, angry, fearful, disgust, surprised, neutral

The combined use of these datasets supports better generalization. Instead of learning only one accent, voice type, or recording environment, the model is exposed to multiple acoustic patterns. This is important for real-time deployment, where users may speak with different accents, emotional intensities, speaking speeds, and microphone conditions.

VI. PREPROCESSING AND FEATURE EXTRACTION

Raw audio signals often contain silence, background noise, amplitude differences, and inconsistent sampling rates. These inconsistencies reduce model performance if they are not handled before feature extraction. The preprocessing stage standardizes the audio and improves signal quality. In the proposed system, preprocessing includes sampling rate normalization, silence removal, noise reduction, framing, windowing, and amplitude normalization.

All audio files are converted to a common sampling rate of 16 kHz. This makes the time-frequency representation consistent and reduces computational complexity. Leading and trailing silent regions are removed because they do not contain emotional information. Noise reduction methods such as spectral filtering are applied to reduce environmental interference. The speech signal is divided into short frames using a Hamming window so that local temporal patterns can be analyzed. Finally, amplitude normalization scales the signal into a consistent range and reduces loudness-related instability during training.

After preprocessing, feature extraction converts the speech waveform into numerical representations suitable for deep learning. Emotional cues are reflected in different aspects of the speech signal. Therefore, the system extracts multiple complementary features rather than relying on a single feature type. MFCC captures speech spectral shape and approximates human auditory perception. ZCR measures the number of times the signal crosses the zero-amplitude axis and can indicate signal sharpness. RMS energy measures intensity variation, which is useful because anger and happiness often have higher energy than sadness or neutral speech. Chroma features capture tonal information, while Mel spectrograms provide two-dimensional time-frequency representations suitable for CNN input.

TABLE II
FEATURES EXTRACTED FROM SPEECH SIGNALS

Feature	Meaning	Importance in SER
MFCC	Short-term spectral representation	Captures vocal tract and frequency patterns
Mel spectrogram	Time-frequency energy map	Primary CNN input for spectral-spatial learning
ZCR	Zero-crossing frequency	Indicates sharpness and noise-like activity
RMS energy	Signal energy level	Represents intensity and loudness variations
Chroma	Pitch class distribution	Captures tonal and harmonic cues

Feature extraction is central to system performance because the model cannot directly interpret human emotion from raw audio without meaningful representation. The combination of spectral, temporal, energy, and tonal descriptors provides a richer view of emotional speech. The Mel spectrogram is especially important because it acts as an image-like representation, enabling CNN layers to learn local patterns of emotional expression.

VII. DATA AUGMENTATION

Emotional speech datasets are often small compared with the number of parameters used in deep learning models. When the dataset is small, a model may memorize training samples rather than learning general emotional patterns. Data augmentation is used to reduce this problem by generating additional variations of existing samples. The augmented samples simulate realistic acoustic changes and improve the model's ability to generalize to unseen speech.

Noise addition introduces Gaussian or environmental noise at different levels. This helps the model remain stable when background noise is present in real use. Pitch shifting changes vocal pitch without changing speech content, allowing the model to handle speaker variation. Time stretching changes speaking speed while preserving pitch, helping the model adapt to fast and slow speech. Spectrogram shifting shifts time-frequency representations to improve robustness and reduce dependence on exact spectral position.

Augmentation is particularly important for real-time SER because real microphone input is rarely as clean as laboratory recordings. A model trained only on clean speech may perform poorly in actual environments. By exposing the model to artificial variations during training, the proposed system becomes more tolerant of noise, pitch variation, and speaking style differences.

TABLE III
DATA AUGMENTATION TECHNIQUES

Technique	Operation	Expected benefit
Noise addition	Adds Gaussian/environmental noise	Improves noise robustness
Pitch shifting	Changes pitch while preserving duration	Simulates speaker and gender variation
Time stretching	Changes speed while preserving pitch	Handles fast and slow speech
Spectrogram shifting	Moves spectrogram along time/frequency axes	Reduces overfitting and improves diversity

VIII. MODEL ARCHITECTURES

Four deep learning models are considered in the study: MLP, CNN, ResNet50, and the proposed CNN+Bi-LSTM hybrid model. The MLP model acts as a baseline. It uses fully connected layers and can learn nonlinear relationships from flattened feature vectors. However, MLP does not explicitly learn spatial patterns from spectrograms or temporal dependencies across speech frames, so its performance is limited for complex SER tasks.

The CNN model improves over MLP by learning local patterns from spectrogram representations. Convolutional filters can detect meaningful spectral structures such as energy bands, frequency changes, and local texture patterns. Pooling layers reduce dimensionality and improve robustness. CNNs are efficient and effective for spectrogram-based learning, but they treat the input mainly as an image and may not fully capture the sequential nature of emotion.

ResNet50 uses residual connections that allow deeper networks to learn complex features without severe vanishing-gradient problems. In Phase I, ResNet50 performed strongly because of its deep feature extraction capability. However, the model also introduced higher computational cost and increased inference latency. This makes it less suitable for lightweight real-time deployment compared with a carefully designed hybrid model.

The proposed CNN+Bi-LSTM model combines the strengths of CNN and temporal sequence learning. CNN layers first extract local spectral-spatial patterns from Mel spectrograms. The resulting feature maps are reshaped into sequential representations and passed to Bi-LSTM layers. The Bi-LSTM processes the sequence in both directions, allowing the model to use previous and future frame context. Dense layers then integrate the learned features and the Softmax layer outputs the final emotion class.

This hybrid approach is well suited for SER because emotional expression depends on both what appears in the spectrum and how those patterns evolve over time. For example, anger may show high energy and rapid changes, sadness may show lower energy and slower movement, and happiness may contain dynamic pitch variation. CNN captures the spectral signature, while Bi-LSTM captures temporal progression.

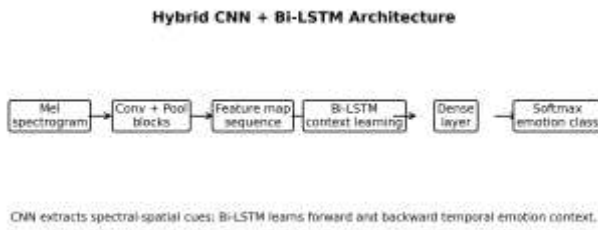


Fig. 3. Conceptual architecture of the hybrid CNN+Bi-LSTM model.

TABLE IV
MODEL ARCHITECTURES COMPARED IN PHASE II

Model	Main strength	Main limitation
MLP	Simple baseline for feature-vector learning	No spatial or temporal modeling
CNN	Learns spectrogram-based spatial features	Limited sequence learning
ResNet50	Deep residual feature extraction	Higher complexity and latency
CNN+Bi-LSTM	Combines spatial and temporal learning	Requires careful training and tuning

IX. REAL-TIME SYSTEM IMPLEMENTATION

The real-time SER system is designed to process live audio input and produce emotion predictions with minimal delay. During real-time operation, the input capture module records speech from a microphone and converts it into digital waveform data. The preprocessing module standardizes the signal, removes silence, reduces noise, and normalizes amplitude. The feature extraction module computes the same features used during training so that the live input has a compatible format.

The inference engine loads the trained CNN+Bi-LSTM model and performs prediction on incoming speech segments. To improve stability, predictions can be smoothed across multiple frames or short utterance windows. The system outputs an emotion label and confidence score. The recognized emotion can be displayed on a user interface or sent to another application for adaptive response.

Latency is a critical requirement for real-time emotion recognition. A system that is accurate but slow may not be suitable for interactive applications. The proposed model achieved an average inference latency of 620 ms, which is lower than MLP, CNN, and ResNet50 in the reported Phase II results. This indicates that the hybrid architecture provides a practical balance between accuracy and response time.

The implementation was conducted using Python with TensorFlow, Keras, Librosa, NumPy, Pandas, Scikit-learn, and OpenCV. The experiments used an Intel Core i7 processor, NVIDIA RTX 3060 GPU, 32 GB RAM, and Windows or Linux operating environment. The training configuration included Adam optimizer, 0.001 learning rate, 32 batch size, 100 epochs, and categorical cross-entropy loss.

TABLE V
TRAINING AND IMPLEMENTATION CONFIGURATION

Component	Configuration
Programming language	Python
Deep learning frameworks	TensorFlow and Keras
Audio processing	Librosa
Optimizer	Adam
Learning rate	0.001
Batch size	32
Epochs	100
Loss function	Categorical cross-entropy

Model: "sequential"

Layer (type)	Output Shape	Param #
conv1d (Conv1D)	(None, 40, 64)	384
activation (Activation)	(None, 40, 64)	0
dropout (Dropout)	(None, 40, 64)	0
max_pooling1d (MaxPooling1D)	(None, 10, 64)	0
conv1d_1 (Conv1D)	(None, 10, 128)	41088
activation_1 (Activation)	(None, 10, 128)	0
dropout_1 (Dropout)	(None, 10, 128)	0
max_pooling1d_1 (MaxPooling1D)	(None, 2, 128)	0
conv1d_2 (Conv1D)	(None, 2, 256)	164096
activation_2 (Activation)	(None, 2, 256)	0
dropout_2 (Dropout)	(None, 2, 256)	0
flatten (Flatten)	(None, 512)	0
dense (Dense)	(None, 8)	4104
activation_3 (Activation)	(None, 8)	0
Total params: 209,672		
Trainable params: 209,672		
Non-trainable params: 0		

Fig. 4. Model summary screenshot from the implementation.

X. EVALUATION METRICS

Model performance is evaluated using accuracy, precision, recall, F1-score, confusion matrix, inference latency, and robustness under noise conditions. Accuracy measures the percentage of total correct predictions. Precision measures the reliability of predicted emotion labels, while recall measures how many actual samples of a class are correctly detected. F1-score combines precision and recall into a balanced measure.

In a multi-class emotion recognition problem, the confusion matrix is especially useful because it shows which emotions are confused with one another. SER models often misclassify emotionally similar categories. For example, fear and surprise can share high pitch and sudden intensity changes, while sadness and neutrality may both contain low energy. The confusion matrix helps identify these overlaps and guides future improvements.

Inference latency measures how quickly the model produces an output after receiving speech input. Low latency is important for live systems. Robustness testing measures how well the model performs under noisy conditions. Together, these metrics provide a more complete evaluation than accuracy alone.

TABLE VI
PERFORMANCE METRICS USED FOR EVALUATION

Metric	Purpose
Accuracy	Overall correct classification rate
Precision	Reliability of predicted emotion classes
Recall	Ability to detect actual emotion classes
F1-score	Balance between precision and recall
Confusion matrix	Class-wise error analysis
Latency	Real-time responsiveness
Noise robustness	Generalization under acoustic interference

XI. RESULTS AND DISCUSSION

The experimental results demonstrate that the proposed CNN+Bi-LSTM model outperforms the baseline models. Table VII presents the Phase II performance comparison. The MLP baseline achieved 88.35% accuracy, while CNN improved the performance to 91.45%. ResNet50 achieved 93.63% accuracy, reflecting the advantage of deeper residual

learning. The proposed CNN+Bi-LSTM model achieved the highest accuracy of 95.42%, along with 94.90% precision, 94.62% recall, and 94.75% F1-score.

TABLE VII
PHASE II PERFORMANCE RESULTS

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)	Latency (ms)
MLP	88.35	87.24	86.70	86.90	950
CNN	91.45	90.84	90.12	90.46	700
ResNet50	93.63	92.80	92.24	92.46	780
CNN+Bi-LSTM	95.42	94.90	94.62	94.75	620

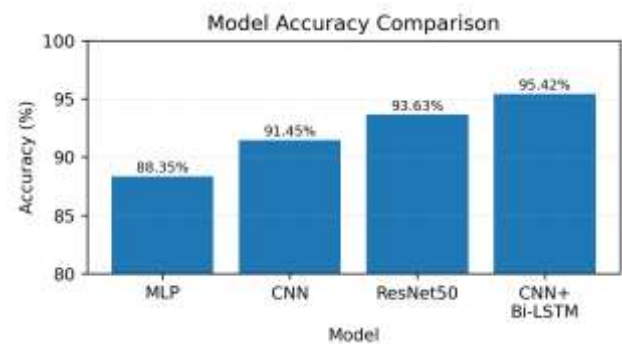


Fig. 5. Accuracy comparison among the evaluated models.

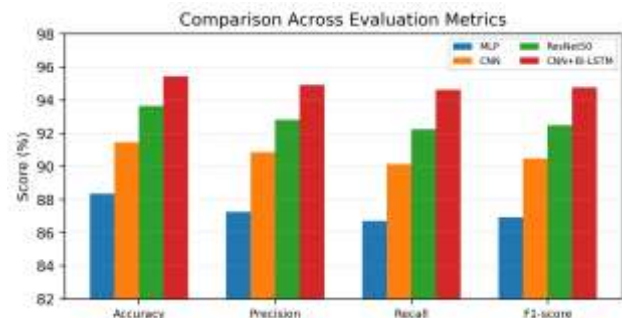


Fig. 6. Comparison of accuracy, precision, recall, and F1-score.

The performance gain of the hybrid model is explained by the complementary behavior of CNN and Bi-LSTM layers. CNN learns useful spectral-spatial features from the Mel spectrogram, while Bi-LSTM interprets the temporal evolution of those features. This allows the system to understand both the acoustic structure and the sequential emotional context. As a result, the hybrid model handles emotional transitions more effectively than standalone CNN and deeper image-based models.

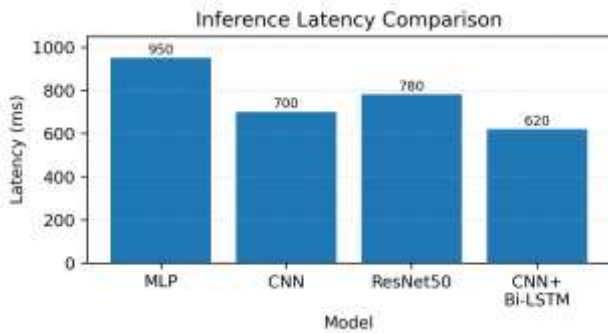


Fig. 7. Inference latency comparison of the evaluated models.

The latency results show that CNN+Bi-LSTM achieved the lowest reported inference latency at 620 ms. ResNet50 produced higher accuracy than CNN but required 780 ms latency, which is slower than the hybrid model. MLP had the highest latency in the reported table, while CNN achieved 700 ms. The CNN+Bi-LSTM model therefore provides the best trade-off between classification accuracy and real-time responsiveness.

The results also show that the proposed model is more suitable for practical deployment than models that only maximize accuracy without considering speed. In interactive systems such as virtual assistants or driver monitoring, a delay of several seconds would reduce usability. The reported latency suggests that the hybrid model can operate in real-time or near-real-time settings when deployed on suitable hardware.

XII. CONFUSION MATRIX AND CLASS-WISE ANALYSIS

The confusion matrix provides deeper insight into how the model classifies different emotions. The thesis implementation includes a confusion matrix for the ResNet50 baseline, showing class-wise prediction patterns among angry, disgust, fear, happy, neutral, and sad emotions. The matrix indicates that anger, fear, and neutral emotions achieve strong diagonal values, while some confusion exists between fear and sad, fear and happy, and neutral and sad classes.

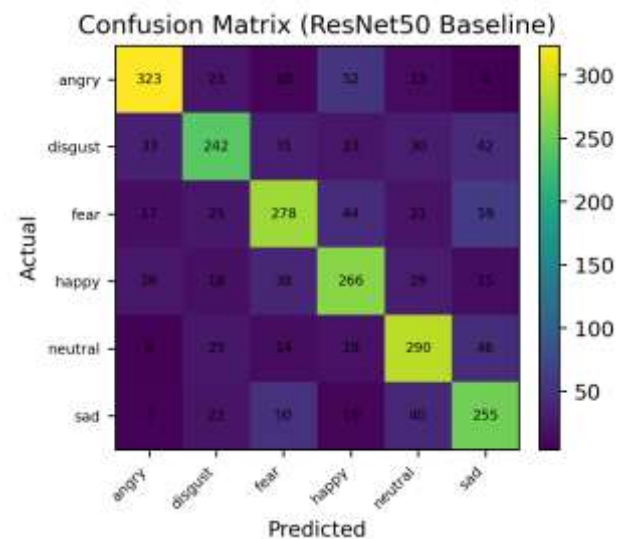


Fig. 8. Confusion matrix visualization based on the implementation output.

Class-wise confusion is expected in SER because some emotions share similar acoustic properties. Fear and surprise or fear and sadness may contain overlapping high-frequency or trembling voice patterns. Sadness and neutrality may share low intensity and slow speech characteristics. Happiness and anger can both contain high energy, although their pitch and rhythm patterns differ. These overlaps explain why confusion matrices are essential for SER evaluation.

The proposed CNN+Bi-LSTM model improves on this challenge by analyzing temporal dynamics rather than relying only on static spectral cues. Emotional speech is not expressed in a single frame. It evolves through pitch movement, rhythm, pauses, and energy changes. Bi-LSTM layers help capture these transitions, making the model more capable of distinguishing emotions with similar local features but different temporal patterns.

```
[12]: [('fear', 'sad', 59),
      ('angry', 'happy', 52),
      ('sad', 'fear', 50),
      ('neutral', 'sad', 46),
      ('fear', 'happy', 44),
      ('disgust', 'sad', 42),
      ('sad', 'neutral', 40),
      ('happy', 'fear', 38),
      ('disgust', 'angry', 33),
      ('disgust', 'fear', 31),
      ('disgust', 'neutral', 30),
      ('happy', 'neutral', 29),
      ('happy', 'angry', 26),
      ('fear', 'disgust', 25),
      ('neutral', 'disgust', 25),
      ('angry', 'disgust', 23),
      ('disgust', 'happy', 23),
      ('sad', 'disgust', 22),
      ('fear', 'neutral', 21),
      ('neutral', 'happy', 19),
      ('happy', 'disgust', 18),
      ('fear', 'angry', 17),
      ('happy', 'sad', 15),
      ('neutral', 'fear', 14),
      ('angry', 'neutral', 13),
      ('angry', 'fear', 10),
      ('sad', 'happy', 10),
      ('sad', 'angry', 7),
      ('neutral', 'angry', 6),
      ('angry', 'sad', 4)]
```

```
[13]: interp.plot_confusion_matrix()
```

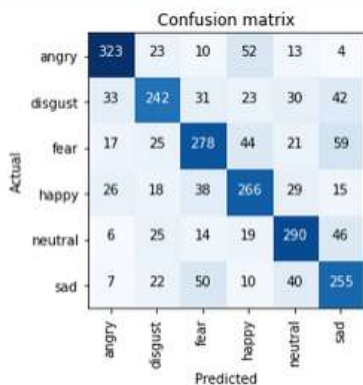


Fig. 9. Confusion matrix screenshot from the original implementation output.

XIII. ROBUSTNESS EVALUATION

Robustness testing evaluates how well the model performs under noisy conditions. Real-time SER systems must handle background noise because practical speech input may come from classrooms, vehicles, offices, hospitals, mobile devices, or public places. The thesis reports accuracy under clean speech and different noise levels. The CNN+Bi-LSTM model achieved 95.42% accuracy on clean speech, 93.85% under 20 dB noise, 91.70% under 15 dB noise, and 88.95% under 10 dB noise.

TABLE VIII
ROBUSTNESS TESTING RESULTS

Noise condition	Accuracy (%)
Clean speech	95.42
20 dB noise	93.85
15 dB noise	91.70
10 dB noise	88.95

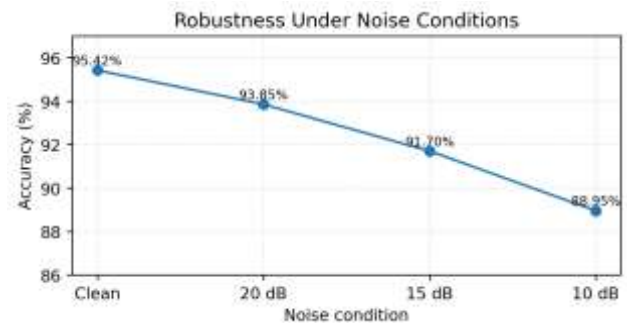


Fig. 10. Accuracy of the proposed model under different noise conditions.

The robustness results show that accuracy decreases as noise increases, which is expected. However, the model maintains strong performance under moderate noise levels. This confirms the value of augmentation techniques such as noise addition and spectrogram shifting. By training the model on varied audio conditions, the system becomes less sensitive to environmental interference and performs more reliably during real-time use.

XIV. PHASE-WISE COMPARATIVE ANALYSIS

Phase I models provided a strong baseline, especially ResNet50, which achieved 93.63% accuracy. However, Phase I focused mainly on spatial feature extraction. It did not fully model temporal sequence information, which is essential for emotional speech. Phase II addressed this issue by introducing the CNN+Bi-LSTM hybrid model. The proposed model improved accuracy to 95.42% while reducing inference latency to 620 ms.

TABLE IX
PHASE I AND PHASE II COMPARISON

Feature	Phase I models	Phase II CNN+Bi-LSTM
Spatial learning	Good	Excellent
Temporal learning	Limited	Strong
Noise robustness	Moderate	High
Real-time adaptability	Moderate	Excellent
Inference latency	Higher	Lower
Generalization	Moderate	Strong

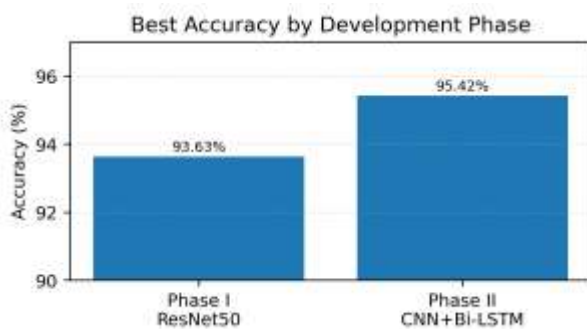


Fig. 11. Phase I and Phase II best-model accuracy comparison.

The phase-wise comparison confirms that a hybrid architecture is more suitable for real-time SER than a standalone spatial learning model. While ResNet50 is powerful, it is computationally heavier and does not explicitly learn bidirectional temporal dependencies. The proposed model improves classification by combining spectral analysis with sequence modeling. This makes it more reliable for live speech input and more adaptable across datasets.

XV. OUTPUT SCREENS AND IMPLEMENTATION EVIDENCE

The uploaded thesis document includes several implementation screenshots from notebook-based experimentation. These screenshots show model training, learning-rate exploration, confusion matrix prediction, model summary, training loss curve, model accuracy curve, and input file structure. Including these screens in the article provides practical evidence that the system was implemented and tested rather than only described theoretically.

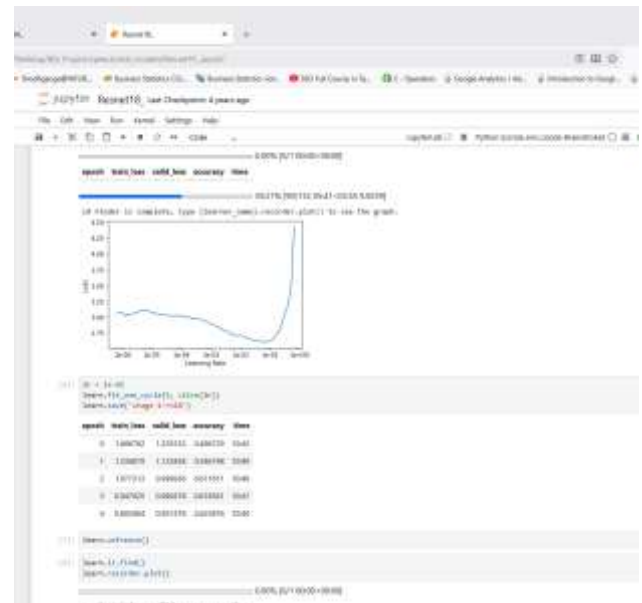


Fig. 12. Training output screenshot from the implementation notebook.

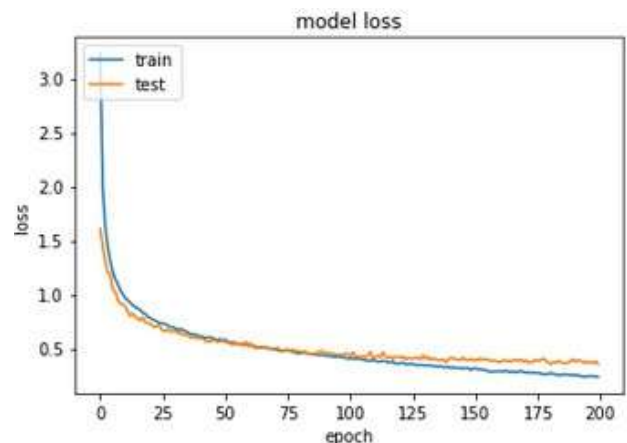


Fig. 13. Model loss curve from the implementation output.

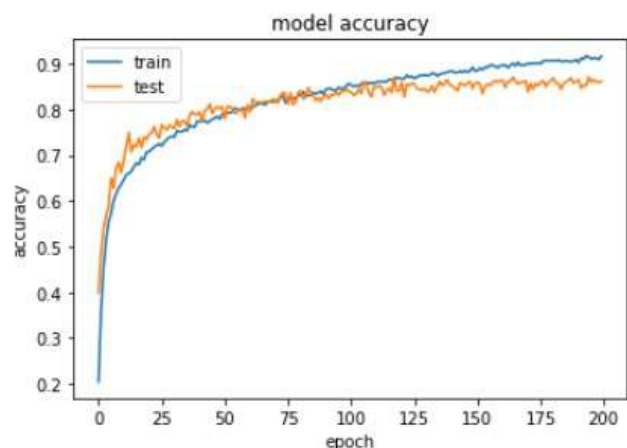


Fig. 14. Model accuracy curve from the implementation output.

The loss and accuracy curves show the training behavior of the implemented model. A decreasing loss curve indicates that the model gradually learned to minimize classification error. The accuracy curve shows improvement over epochs and indicates that the model became better at identifying emotion classes. Stable

validation performance is important because it shows that the model is not merely memorizing training data.

XVI. APPLICATIONS OF THE PROPOSED SYSTEM

The proposed SER system can support several real-world applications. In healthcare, it may assist in monitoring emotional stress, anxiety, or depressive tendencies during telemedicine sessions. It cannot replace medical diagnosis, but it can provide useful emotional indicators for clinicians and caregivers. In customer support, the system can detect anger or frustration and alert supervisors when a user requires urgent attention.

In online education, SER can help identify whether students are bored, confused, stressed, or engaged. This information can be used to improve teaching strategies and adaptive learning systems. In intelligent vehicles, emotion recognition can support driver safety by detecting anger, fatigue, or stress. In virtual assistants and social robots, emotion awareness can make interactions more natural, empathetic, and personalized.

The system can also be integrated into accessibility tools. Users with limited ability to express emotions through facial expressions may benefit from speech-based emotion analysis. In smart homes, emotion-aware interfaces may adjust responses based on the user's vocal state. These applications demonstrate that SER is not only a technical classification problem but also an important step toward human-centered artificial intelligence.

XVII. LIMITATIONS

Although the proposed system achieved strong results, several limitations remain. First, SER performance depends on the quality and diversity of training data. Benchmark datasets are often recorded in controlled environments, while real-world speech may contain spontaneous expressions, overlapping speech, microphone distortion, and uncontrolled noise. Therefore, additional testing on real-world data is necessary before deployment.

Second, emotional categories are often simplified into labels such as happy, sad, angry, fear, disgust, surprise, and neutral. Human emotions are more complex and may be mixed or subtle. A speaker may sound tired and sad at the same time, or angry and frustrated with different intensity levels. Future systems should consider emotion intensity and multi-label emotion recognition.

Third, the model may be affected by language, culture, accent, gender, and speaker personality. Emotional expression varies across people and communities. A model trained on a limited set of speakers may not generalize equally well to all users. This raises fairness and reliability concerns, especially if SER is used in high-impact applications.

Fourth, deep learning models can be difficult to interpret. A high-performing model may not clearly explain why it predicted a specific emotion. Explainable AI techniques are necessary to make the system more transparent. Finally, privacy and consent are important. Speech recordings may contain personal and sensitive information, so real-time SER systems must use secure and ethical data handling practices.

XVIII. CONCLUSION

This paper presented an expanded IEEE-style article on real-time Speech Emotion Recognition using deep learning techniques. The proposed system uses TESS, EmoDB, and RAVDESS datasets, extracts multiple features including MFCC, Mel spectrogram, ZCR, RMS, and Chroma, and applies augmentation methods such as noise addition, pitch shifting, time stretching, and spectrogram shifting. The system compares MLP, CNN, ResNet50, and a hybrid CNN+Bi-LSTM model.

The experimental results show that CNN+Bi-LSTM achieved the best performance with 95.42% accuracy, 94.90% precision, 94.62% recall, 94.75% F1-score, and 620 ms inference latency. The hybrid model outperformed baseline architectures because it combined CNN-based spectral-spatial learning with Bi-LSTM-based temporal sequence learning. Robustness testing also confirmed that the model maintains good performance under noise conditions.

The proposed SER framework is suitable for real-time applications in healthcare, education, customer service, driver monitoring, virtual assistants, and social robotics. The study confirms that temporal modeling is essential for accurate emotion recognition from speech and that data augmentation improves practical robustness. Overall, the CNN+Bi-LSTM framework provides a strong foundation for intelligent emotion-aware systems.

XIX. FUTURE WORK

Future work can focus on multilingual and cross-cultural SER. The current study uses English and German benchmark datasets, but global deployment requires support for many languages, accents, and cultural speech patterns. Cross-lingual transfer learning

may help models adapt to new languages with limited labeled data.

Multimodal emotion recognition is another important direction. Human emotions are expressed not only through speech but also through facial expressions, gestures, eye movement, and physiological signals. Combining speech with visual and physiological information may improve accuracy, especially for subtle or mixed emotions.

Future research can also explore transformer-based models, attention mechanisms, and self-supervised speech representations. These methods may capture long-range dependencies and improve performance on spontaneous speech. Lightweight deployment techniques such as pruning, quantization, knowledge distillation, and edge optimization can reduce model size and make SER suitable for mobile and embedded devices.

Privacy-preserving approaches such as federated learning should also be investigated. SER systems may process sensitive voice data, so training models without centralizing user recordings can improve privacy. Explainable AI methods such as attention visualization and feature attribution should be integrated to make predictions more understandable for researchers, developers, and end users.

REFERENCES

- [1] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR), 2017, pp. 1251-1258.
- [2] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR), 2016, pp. 770-778.
- [3] X. Li, X. Zhao, and X. Li, "Speech emotion recognition using ResNet and attention mechanism," Applied Sciences, vol. 11, no. 8, p. 3529, 2021.
- [4] Y. Zhang, J. Dang, and M. Li, "Deep residual networks for speech emotion recognition in noisy environments," IEEE Access, vol. 10, pp. 51324-51335, 2022.
- [5] Q. Mao, M. Dong, Z. Huang, and Y. Zhan, "Learning salient features for speech emotion recognition using convolutional neural networks," IEEE Transactions on Multimedia, vol. 21, no. 3, pp. 516-525, 2019.
- [6] T. L. Nwe, S. W. Foo, and L. C. De Silva, "Speech emotion recognition using hierarchical deep residual networks," Neural Processing Letters, vol. 51, no. 2, pp. 1503-1519, 2020.
- [7] P. Verma and R. Singh, "Comparative study of deep learning architectures for speech emotion recognition," International Journal of Speech Technology, vol. 24, no. 2, pp. 487-496, 2021.
- [8] P. Tzirakis, G. Trigeorgis, M. A. Nicolaou, B. W. Schuller, and S. Zafeiriou, "End-to-end speech emotion recognition using deep neural networks," IEEE Journal of Selected Topics in Signal Processing, vol. 11, no. 8, pp. 1301-1309, 2017.
- [9] O. Abdel-Hamid, A. R. Mohamed, H. Jiang, L. Deng, G. Penn, and D. Yu, "Convolutional neural networks for speech recognition," IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 22, no. 10, pp. 1533-1545, 2014.
- [10] Y. Zeng, Y. Yin, and J. Yang, "Hybrid CNN-BiLSTM model for speech emotion recognition based on spectrogram and temporal context," Sensors, vol. 23, no. 6, p. 2887, 2023.
- [11] S. Latif, R. Rana, S. Younis, J. Qadir, and B. Schuller, "Survey of deep learning for speech emotion recognition," IEEE Transactions on Affective Computing, vol. 12, no. 3, pp. 546-565, 2020.
- [12] H. M. Fayek, M. Lech, and L. Cavedon, "Evaluating deep learning architectures for speech emotion recognition," Neural Networks, vol. 92, pp. 60-68, 2017.
- [13] Y. Kim and H. Kim, "Deep residual CNN with LSTM layers for robust speech emotion recognition in real-world conditions," IEEE Access, vol. 10, pp. 89102-89114, 2022.
- [14] S. Tripathi, A. Kumar, and R. Singh, "Speech emotion recognition using hybrid deep learning model," Multimedia Tools and Applications, vol. 80, no. 22, pp. 34115-34131, 2021.
- [15] X. Huang, D. Wang, and J. Li, "Temporal feature learning with CNN and Bi-LSTM for real-time speech emotion recognition," Expert Systems with Applications, vol. 224, p. 120007, 2023.