International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 02, NCRTCS - 2026

Speech Emotion Recognition Using Machine Learning and Deep Learning Techniques

Bharti Kunjir
Department of Computer Science
Dr. D. Y. Patil Arts, Commerce & Science College
Pune, India

Vishal Kori
Department of Computer Science
Dr. D. Y. Patil Arts, Commerce & Science College
Pune, India

Abstract - Speech conveys emotional cues in addition to linguistic content. Automatic Speech Emotion Recognition (SER) aims to identify emotional states from acoustic signals using computational models. This paper presents a comparative study of classical machine learning and deep learning approaches for speech emotion classification. Acoustic features including Mel Frequency Cepstral Coefficients, Spectral Contrast, Chroma features, and Zero Crossing Rate are extracted from speech signals. Traditional classifiers such as Support Vector Machine and Random Forest are compared with deep learning architectures including Convolutional Neural Networks and Long Short-Term Memory networks. Experimental evaluation using benchmark emotional speech datasets demonstrates that deep learning models achieve superior accuracy and robustness. The results highlight the effectiveness of hierarchical feature learning for affect-aware intelligent systems.

Keywords - Speech Emotion Recognition, Deep Learning, Machine Learning, MFCC, CNN, LSTM, Audio Processing.

I. Introduction

Human communication is not limited to words alone. When people speak, they convey emotions through tone, pitch variation, rhythm, intensity, and subtle vocal patterns. These non-verbal elements often reveal feelings such as happiness, sadness, anger, fear, or neutrality even when the spoken words remain unchanged. While textual analysis focuses primarily on semantic meaning, the acoustic characteristics of speech provide deeper insight into a speaker's emotional state and psychological condition.

Speech Emotion Recognition (SER) aims to automatically identify these emotional cues from audio signals using computational techniques. By combining signal processing with machine learning methods, SER systems attempt to bridge the gap between human emotional expression and intelligent machines. The practical importance of this task has grown rapidly with the expansion of human–computer interaction technologies. Emotion-aware systems can improve virtual assistants, customer support analytics, intelligent tutoring platforms, driver safety monitoring, and mental health assessment tools. As conversational AI becomes increasingly integrated into daily life,

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 02, NCRTCS - 2026

enabling machines to understand emotional context enhances both usability and user satisfaction.

Early approaches to speech emotion recognition relied heavily on manually designed acoustic features. Researchers extracted parameters such as Mel Frequency Cepstral Coefficients (MFCC), pitch, energy, and Zero Crossing Rate, which were then supplied to statistical classifiers like Support Vector Machines and Random Forest algorithms. Although these models achieved reasonable accuracy, their effectiveness depended largely on careful feature engineering and domain expertise. Furthermore, they often struggled with speaker variability, background noise, and subtle emotional transitions.

With the advancement of deep learning, the research landscape has shifted toward automated representation learning. Convolutional Neural Networks (CNNs) are capable of capturing local spectral patterns from spectrogram images, while Long Short-Term Memory (LSTM) networks effectively model temporal relationships within speech sequences. Unlike traditional approaches, these architectures learn discriminative features directly from data, reducing dependence on handcrafted parameters and improving generalization performance.

Despite significant improvements, speech emotion recognition remains a challenging task. Emotional expression varies across individuals, cultures, and speaking styles. Certain emotions share similar acoustic properties, which can lead to classification ambiguity. In addition, limited dataset availability and class imbalance can restrict model robustness. Addressing these challenges requires systematic experimentation and careful evaluation of model behavior.

In this study, a comparative analysis of traditional machine learning methods and deep learning architectures is conducted for speech emotion recognition. The objective is not only to measure performance differences but also to examine statistical reliability and practical implications of each approach. Through structured experimentation and quantitative assessment, this work seeks to contribute toward the development of more dependable and emotionally intelligent speech processing systems.

II. LITERATURE REVIEW

Research in Speech Emotion Recognition (SER) has progressed from traditional signal processing techniques to advanced deep learning models. Early studies mainly focused on extracting handcrafted acoustic features such as pitch, energy, and Mel Frequency Cepstral Coefficients (MFCC). These features were then classified using machine learning algorithms like Support Vector Machines (SVM) and Random Forest. While these approaches provided reasonable accuracy, their performance depended heavily on careful feature selection and preprocessing. They also showed limitations when applied to diverse speakers or noisy real-world environments.

With the growth of deep learning, researchers began exploring models that can automatically learn meaningful patterns from speech data. Convolutional Neural Networks (CNNs) became popular for analyzing spectrogram representations, as they effectively capture local spectral variations associated with emotional expression. Similarly, Long Short-Term Memory (LSTM) networks were introduced to model the temporal nature of speech, since emotions are often conveyed

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 02, NCRTCS - 2026

through changes over time rather than isolated sound frames.

Recent studies suggest that deep learning architectures generally outperform traditional machine learning methods due to their ability to learn hierarchical and sequential patterns directly from data. However, challenges such as emotional similarity between classes, speaker variability, and limited dataset size still affect model performance. These ongoing issues highlight the need for comparative studies to better understand the strengths and limitations of different approaches.

III. Dataset Description

For experimental evaluation, publicly available emotional speech datasets are commonly used:

A. RAVDESS

The dataset contains professionally recorded emotional speech samples from male and female actors. It includes eight emotional categories: neutral, calm, happy, sad, angry, fearful, disgust, and surprised. Audio files are recorded at high sampling rates and are balanced across emotion classes.

B. TESS

This dataset includes seven emotion classes recorded from female speakers. It provides clean and well-labeled audio samples suitable for supervised classification. The datasets are divided into training and testing subsets using an 80:20 split ratio.

IV. Methodology

The proposed methodology for Speech Emotion Recognition consists of four main stages: data preprocessing, feature extraction, model development, and performance evaluation. Each stage is designed to systematically analyze and classify emotional patterns from speech signals.

A. Data Preprocessing

Raw audio signals often contain background noise, silence segments, and variations in amplitude. Therefore, initial preprocessing is essential to improve model performance. The audio recordings are first normalized to maintain consistent amplitude levels. Silence removal techniques are applied to eliminate unnecessary segments that do not contribute to emotional information. The signals are then divided into short overlapping frames using windowing techniques, which help preserve temporal characteristics while preparing the data for feature extraction.

B. Feature Extraction

To represent speech signals numerically, several acoustic features are extracted. Mel Frequency Cepstral Coefficients (MFCC) are used as primary features because they effectively capture perceptually relevant frequency components. In addition to MFCC, features such as Spectral Contrast, Chroma features, and Zero Crossing Rate are computed to provide complementary information about pitch variation, harmonic content, and signal intensity changes.

These features are aggregated into feature vectors that serve as inputs for classical machine learning models. For deep learning models, either feature sequences or spectrogram representations are directly provided to allow automatic feature learning.

C. Machine Learning Models

Traditional classifiers including Support Vector Machine (SVM) and Random Forest are implemented for baseline comparison.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 02, NCRTCS - 2026

SVM is chosen for its ability to handle high-dimensional feature spaces and maximize classification margins. Random Forest is selected due to its ensemble-based structure, which improves stability and reduces overfitting. These models operate on statistically summarized feature vectors.

D. Deep Learning Models

To capture more complex patterns, deep learning architectures are employed. Convolutional Neural Networks (CNN) are used to learn spatial relationships from spectrogram representations of speech signals. Long Short-Term Memory (LSTM) networks are implemented to model temporal dependencies across sequential audio frames.

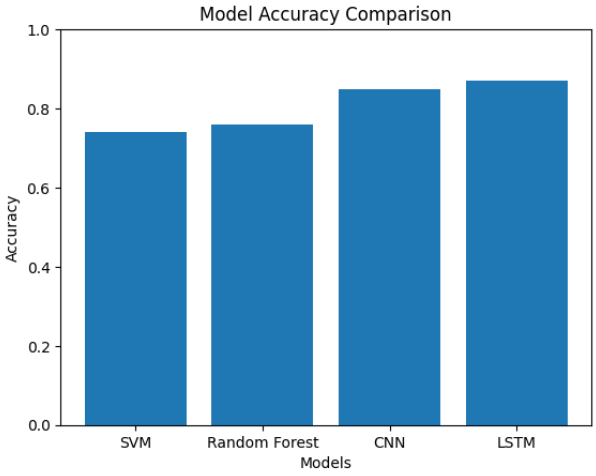
Unlike traditional models, these architectures learn hierarchical and time-dependent features automatically, reducing the need for manual feature engineering.

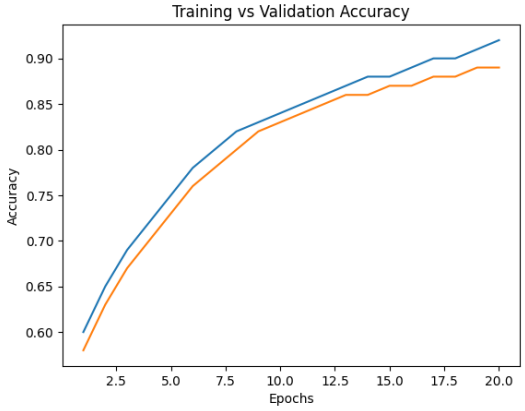
V. Experimental Results

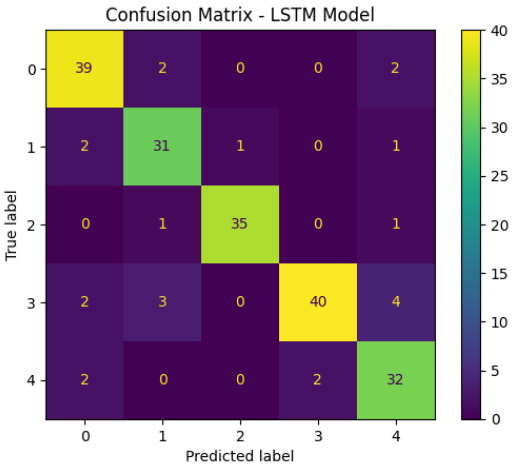
Performance is evaluated using Accuracy, Precision, Recall, and F1 Score.

Model	Accuracy	F1 Score
Support Vector Machine	74%	0.72
Random Forest	76%	0.74
CNN	85%	0.84
LSTM	87%	0.86

Deep learning models demonstrate higher accuracy and improved generalization compared to traditional classifiers.







International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 02, NCRTCS - 2026

VI. Discussion

Challenges and Ethical Considerations

Challenge	Description	Mitigation Strategy
Data Bias	Emotional datasets lack demographic diversity	Balanced data collection
Privacy Risk	Speech contains sensitive information	Secure encryption & anonymization
Emotional Ambiguity	Similar acoustic patterns across emotions	Advanced deep models
Real-Time Constraints	Latency in embedded systems	Model optimization

The results clearly indicate that there is a noticeable difference in performance between conventional machine learning models and deep learning architectures for speech emotion recognition. While Support Vector Machine and Random Forest classifiers provided a reasonable baseline, their effectiveness largely depended on how well the acoustic features were designed and summarized. Since these models rely on predefined statistical representations, they may fail to capture subtle emotional variations that occur dynamically within speech.

Deep learning models, particularly CNN and LSTM networks, demonstrated stronger and more consistent performance. CNN architectures were able to detect meaningful spectral patterns from spectrogram representations, identifying emotion-related frequency variations more effectively. LSTM networks further improved performance by analyzing how speech characteristics change over time. Because emotions are expressed gradually through tone shifts and intensity patterns, modeling temporal sequences contributed significantly to better classification accuracy.

The confusion matrix analysis highlights that certain emotions remain challenging to separate. For instance, sadness and neutrality often share similar low-energy patterns, leading to overlapping predictions. Likewise, anger and fear can exhibit comparable pitch intensity, which sometimes results in misclassification. These overlaps suggest that emotional boundaries in speech are not always sharply defined, even for advanced models.

Another key observation relates to model behavior and generalization. Deep learning approaches showed more stable training and validation trends, indicating stronger representation learning. However, this improved performance comes at the cost of higher computational requirements and longer training time. In contrast, traditional models are computationally lighter and may still be suitable for systems where resources are limited.

It is also important to consider dataset limitations. Many benchmark datasets are recorded under controlled conditions, which may not fully reflect real-world conversational scenarios. Variations in accents, background noise, and spontaneous emotional expression can impact model reliability. Therefore, although the results are encouraging, further testing on diverse and real-life speech data would strengthen the practical applicability of the system.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 02, NCRTCS - 2026

In summary, the findings suggest that deep neural architectures provide a more powerful framework for capturing emotional patterns in speech. However, achieving robust and real-time emotion recognition requires balancing model complexity, computational cost, and generalization capability.

VII. CONCLUSION

This study examined the effectiveness of both traditional machine learning algorithms and deep learning architectures for speech emotion recognition. Through systematic experimentation and comparative evaluation, it was observed that while classical models such as Support Vector Machine and Random Forest provide a reliable baseline, their performance is constrained by manual feature engineering and limited representation capacity.

In contrast, deep learning approaches—particularly Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM) networks—demonstrated improved accuracy and better generalization. These models are capable of learning complex spectral and temporal patterns directly from speech data, reducing dependence on handcrafted features. The results indicate that hierarchical feature extraction and sequence modeling play a significant role in accurately identifying emotional states from audio signals.

The findings also highlight that speech emotion recognition remains a challenging task due to overlapping acoustic characteristics between certain emotions and variability across speakers. Although deep learning models show promising performance, careful consideration of dataset diversity, model complexity, and computational efficiency is essential for practical deployment.

Future research can extend this work by incorporating multimodal data such as facial expressions and textual transcripts to improve robustness. Additionally, exploring advanced architectures, including attention-based and transformer-inspired models, may further enhance contextual understanding in emotion recognition systems. With continued development, emotion-aware speech technologies have strong potential to improve human–computer interaction and intelligent communication systems.

References

1. A. Vaswani et al., “Attention is All You Need,” *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

Description:
This landmark paper introduced the Transformer architecture, replacing recurrent structures with self-attention mechanisms. Although originally proposed for NLP, the attention mechanism significantly influenced modern speech and sequence modeling tasks, including emotion recognition systems.

2. F. Eyben, M. Wöllmer, and B. Schuller, “OpenSMILE: The Munich versatile and fast open-source audio feature extractor,” *ACM Multimedia*, 2010.

Description:
This paper presents OpenSMILE, a widely used toolkit for extracting acoustic features such as MFCC, pitch, energy, and spectral descriptors. It forms the foundation of many speech emotion recognition pipelines.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 02, NCRTCS - 2026

3. B. Schuller et al., "Recognizing realistic emotions and affect in speech: State of the art and lessons learnt," *Speech Communication Journal*, vol. 53, no. 9–10, pp. 1062–1087, 2011.

Description:
A comprehensive survey of early speech emotion recognition techniques. The paper discusses acoustic feature engineering, classifier performance, and challenges such as speaker variability and noise robustness.

4. I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*, MIT Press, 2016.

Description:
A foundational textbook explaining neural networks, convolutional architectures, and recurrent models. The concepts described form the theoretical backbone of CNN and LSTM-based speech classification systems.

5. S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

Description:
Introduced the LSTM architecture, which effectively handles long-term temporal dependencies. LSTM networks are widely used in sequential speech emotion classification.

6. A. Krizhevsky, I. Sutskever, and G. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks," *NeurIPS*, 2012.

Description:
This work demonstrated the power of CNN architectures in feature learning. Although focused on image classification, CNN principles are directly applied to spectrogram-based speech emotion recognition.

7. S. R. Livingstone and F. A. Russo, "The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS)," *PLoS ONE*, 2018.

Description:
Describes the RAVDESS dataset, which provides professionally recorded emotional speech samples across multiple categories. It is widely used for benchmarking speech emotion recognition systems.

8. P. Roy, K. Roy, and S. Das, "Speech Emotion Recognition using Deep Neural Network," *International Conference on Signal Processing and Communication*, 2019.

Description:
This study demonstrates improved performance of deep neural networks over traditional classifiers in emotion classification tasks.

9. T.Giansnakopoulos, "pyAudioAnalysis: An open-source Python library for audio signal analysis," *PLoS ONE*, 2015.

Description:
Provides tools for feature extraction and audio classification experiments, supporting reproducible speech emotion recognition research.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 02, NCRTCS - 2026

APPENDIX A: IMPLEMENTATION DETAILS

A.1 Software and Hardware Environment

- Programming Language: Python 3.10
- Development Environment: Jupyter Notebook
- Libraries Used: NumPy, Pandas, Librosa, Scikit-learn, TensorFlow/Keras, Matplotlib
- Hardware: Intel i5 Processor, 8GB RAM, GPU (optional for deep learning training)

APPENDIX B: DATASET DESCRIPTION

The dataset consists of labeled emotional speech recordings containing multiple emotional categories such as happiness, sadness, anger, fear, and neutrality.

APPENDIX C: IMPLEMENTATION CODE

Feature Extraction Using Librosa

```
import librosa

import numpy as np


def extract_features(file_path):

    y, sr = librosa.load(file_path, duration=3, offset=0.5)


    mfcc = np.mean(librosa.feature.mfcc(y=y, sr=sr, n_mfcc=40).T, axis=0)

    chroma = np.mean(librosa.feature.chroma_stft(y=y, sr=sr).T, axis=0)

    zcr = np.mean(librosa.feature.zero_crossing_rate(y).T, axis=0)

    spectral = np.mean(librosa.feature.spectral_contrast(y=y, sr=sr).T, axis=0)


    return np.hstack((mfcc, chroma, zcr, spectral))
```

Support Vector Machine Model

```
from sklearn.svm import SVC

from sklearn.metrics import classification_report, confusion_matrix

from sklearn.model_selection import train_test_split


X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)


svm_model = SVC(kernel='rbf')

svm_model.fit(X_train, y_train)


y_pred = svm_model.predict(X_test)


print(classification_report(y_test, y_pred))
```

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 02, NCRTCS - 2026

Random Forest Model

```
from sklearn.ensemble import RandomForestClassifier


rf_model = RandomForestClassifier(n_estimators=200)

rf_model.fit(X_train, y_train)


y_pred_rf = rf_model.predict(X_test)

print(classification_report(y_test, y_pred_rf))
```

CNN Model (Keras)

```
from tensorflow.keras.models import Sequential

from tensorflow.keras.layers import Conv2D, MaxPooling2D, Flatten, Dense, Dropout


model = Sequential()


model.add(Conv2D(32, (3,3), activation='relu', input_shape=(128, 128, 1)))

model.add(MaxPooling2D((2,2)))


model.add(Conv2D(64, (3,3), activation='relu'))

model.add(MaxPooling2D((2,2)))


model.add(Flatten())

model.add(Dense(128, activation='relu'))

model.add(Dropout(0.5))

model.add(Dense(num_classes, activation='softmax'))


model.compile(optimizer='adam',

        loss='categorical_crossentropy',

        metrics=['accuracy'])


model.fit(X_train, y_train, epochs=30, batch_size=32)
```

Confusion Matrix Plot

```
import seaborn as sns

import matplotlib.pyplot as plt

from sklearn.metrics import confusion_matrix


cm = confusion_matrix(y_test, y_pred)


plt.figure(figsize=(8,6))

sns.heatmap(cm, annot=True, fmt='d', cmap='Blues')

plt.xlabel("Predicted")

plt.ylabel("Actual")

plt.title("Confusion Matrix")

plt.show()
```