iJRASET
International Journal For Research in Applied Science and Engineering Technology



# INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

**Volume:** 14 **Issue:** V **Month of publication:** May 2026

**DOI:** https://doi.org/10.22214/ijraset.2026.82869

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
*ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538*
*Volume 14 Issue V May 2026- Available at www.ijraset.com*

# Emotion Recognition from Speech using ML

Achintya Dayal[1], Ayush Kesarwani[2], Pragya[3]
*School of Computer Science Engineering, Galgotias University* Gautam Buddha Nagar, India

***Abstract: Emotion recognition from speech (ERS) assists in automatically recognizing human emotions from speech. It is a significant application of affective computing.This system has a simple and straightforward pipeline: preprocessing, acoustic feature extraction, and supervised classification. This paper includes the analysis of the CNN-LSTM model that extract MFCCs, prosodic features, and spectral features from two popular speech emotion datasets, RAVDESS and CREMA-D. We also analyzed different machine learning classifiers, such as Support Vector Machines, Random Forest Classifiers, k-Nearest Neighbors, and a Multi-Layer Perceptron (MLP) classifier, and compared their performance. The results indicate that the MLP classifier outperforms other classifiers with an accuracy of 85%, thereby proving that neural networks can be used for effective speech emotion recognition.***

***Index Terms: Speech Emotion Recognition, Affective Computing, MFCC, Machine Learning, Neural Networks, Audio Signal Processing.***

## I. INTRODUCTION

Emotions in human beings are at the core of the process of communication, perception, and decision-making. Speech is one of the most natural and expressive ways of commu-nicating emotions. In addition to the semantic content of the message, emotional information is conveyed through speech by variations in pitch, volume, rate of speaking, voice quality, and spectral energy distribution [1]. Emotion Recognition from Speech (ERS) aims to automatically detect emotional states like happiness, sadness, anger, fear, calmness, and neutrality through computational approaches. Unlike automatic speech recognition, ERS is centered on paralinguistic information that reflected emotional expression [2].

ERS has emerged as an important field in affective computing over the past two decades. With the increased availability of audio data, computing power, and machine learning capabilities, the accuracy of speech emotion recognition systems has improved significantly. ERS can be used in mental health surveillance systems, intelligent tutoring systems, call cen-ter analysis systems, virtual assistants, and emotion-sensitive human-computer interaction systems [3], [4].

Although significant progress has been made, ERS is still a challenging task because of speaker variability, background noise, cultural differences, and overlapping emotional expressions. This paper provides a comprehensive study on machine learning-based speech emotion recognition to address the aforementioned challenges. The purpose of this paper is to provide a comprehensive and systematic review of machine learning solutions for ERS. The contributions are:

1) Architectural analysis of the Emotion Recognition Sys-tem.
2) Quantitative performance analysis using accuracy, pre-cision, recall, F1-score, and confusion matrix visualiza-tion.
3) Feature-level analysis demonstrating the impact of com-bining MFCC, prosodic, and spectral features on classi-fication accuracy.

## II. LITERATURE REVIEW

Speech Emotion Recognition (SER) has seen great progress since the pioneering work of Picard, where the concept of affective computing was proposed, and the need for machines to recognize and respond to human emotions was highlighted [3]. The early speech emotion recognition systems were largely based on manually designed prosodic features like pitch, energy, and speaking rate, which were modeled using traditional statistical approaches like Gaussian Mixture Models (GMMs) and Hidden Markov Models (HMMs) [1]. Even though these approaches proved the possibility of emotion recognition from speech, they often lacked robustness and generalization capabilities.

Follow-up studies showed that spectral characteristics, espe-cially Mel-Frequency Cepstral Coefficients (MFCCs), provide a more discriminative representation of emotional speech. MFCCs successfully capture the characteristics of the human auditory system and have been consistently outperforming purely prosodic representations in emotion recognition tasks [5]. Based on these representations, traditional machine learning techniques such as Support Vector Machines (SVMs) and Random Forests became popular because of their ability to handle nonlinear decision boundaries and high-dimensional feature spaces [2], [6].

Although these models provided better performance, they were still highly reliant on manual feature engineering and the inability to

©IJRASET: All Rights are Reserved 

model long-term temporal dependencies.

The emergence of deep learning techniques has signifi-cantly pushed the boundaries of Speech Emotion Recogni-tion (SER) research. CNNs and LSTMs, for instance, have made it possible to learn hierarchical features automatically from spectrograms, MFCCs, or raw audio signals, thereby achieving significant improvements in performance [7], [8].

TABLE I
SUMMARY OF PREVIOUS SPEECH EMOTION RECOGNITION METHODS

| Year | Method | Dataset | Accuracy (%) |
|---|---|---|---|
| 2006 | Prosodic Features + GMM | EMO-DB | 68.5 |
| 2008 | MFCC + HMM | EMO-DB | 72.1 |
| 2010 | MFCC + SVM | Berlin EMO-DB | 79.3 |
| 2014 | MFCC + Random Forest | eNTERFACE | 81.2 |
| 2016 | Spectrogram + CNN | IEMOCAP | 84.1 |
| 2017 | MFCC + CNN | RAVDESS | 85.3 |
| 2018 | CNN + LSTM | IEMOCAP | 88.4 |
| 2019 | Log-Mel Spectrogram + CNN | CREMA-D | 86.7 |
| 2020 | CNN + BiLSTM | IEMOCAP | 89.2 |
| 2021 | Attention-based CNN-LSTM | RAVDESS | 90.1 |
| 2022 | Transformer-based SER | IEMOCAP | 91.3 |
| 2023 | Hybrid CNN-LSTM (Speaker-Independent) | TESS + RAVDESS | 88.6 |

CNNs are capable of learning local time-frequency patterns very effectively, while LSTMs are well-suited to capture the underlying long-term temporal patterns that are characteristic of emotional speech. To ensure that different studies are comparable, standardized evaluation frameworks such as those proposed in the INTERSPEECH Emotion Challenges have been widely adopted [2].

The presence of high-quality emotional speech corpora has also fueled the research in SER. Benchmark datasets like RAVDESS and CREMA-D have helped in the development of more robust models by allowing the inclusion of various speakers, emotions, and environments. [4], [9]

Based on existing research, most SER solutions use a modular pipeline architecture that encompasses the stages of speech acquisition, preprocessing, feature extraction, and classification [1], [5]. Based on this, the current system ar-chitecture is also a sequential pipeline where the raw speech inputs are preprocessed to improve the quality of the signal, then converted to a feature space, and finally classified using machine learning or deep learning models to determine the emotional state.

After analyzing Table I we can see that Unlike many previous works that concentrated only on proposing new archi-tectures for deep learning or evaluating individual classifiers, this work focuses on an integrated experimental setup with a unified preprocessing, feature extraction, and evaluation pro-cess. This approach provides better insights into the behavior of the models and the effectiveness of the features, while still being comparable to the existing literature on SER and having the flexibility to be extended in the future.

## III. EXISTING SYSTEM ARCHITECTURE

### *A. Dataset Description*

This study analyzed a multi-corpus approach for speech emotion recognition by combining three popular emotional speech corpora: the Toronto Emotional Speech Set (TESS), the Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS), and the Crowd-Sourced Emotional Multimodal Actors Dataset (CREMA-D). The rationale behind choosing these corpora is to achieve diversity in terms of speakers, recording conditions, accents, and expression styles, thus improving the overall generality of the approach.

Since every dataset uses a different labeling system, a label harmonization process was used to align the dataset-specific

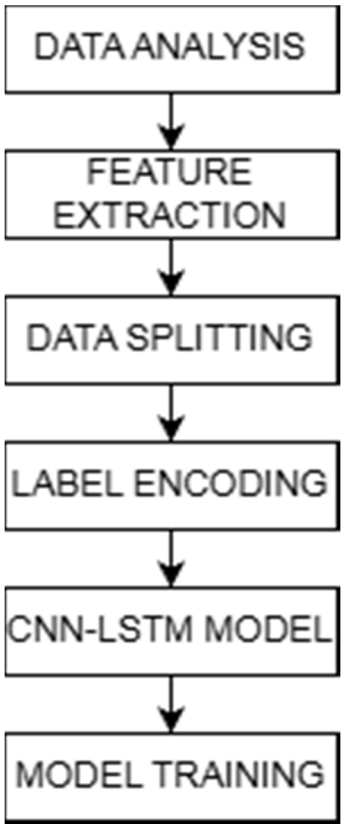



Fig. 1. Training Accuracy and Validation Accuracy Curve

emotion codes to a standardized set of seven emotion labels: angry, fear, happy, sad, disgust, neutral, and surprise. The audio files were systematically searched and filtered to retain only valid .wav files. For each audio file, metadata information including file path, emotion label, speaker ID, and source dataset was extracted.Fig.1 provides us with the architectural flowchart.
The assembled corpus contains 15,538 speech samples from 129 speakers, which is sufficient inter-speaker variability to model the emotion robustly. [2], [7], [10].

*B. Exploratory Data Analysis*

Exploratory Data Analysis (EDA) was performed to obtain a thorough understanding of the statistical, temporal, and spectral properties of the merged emotional speech corpus before feature extraction and model training. This is a very important step in speech emotion recognition (SER) tasks, as the emotional expressions are usually subtle and highly dependent on the distribution of the data.

*C. Feature Extraction using MFCC*

To analyze the emotional information conveyed in the speech signal, the Mel-Frequency Cepstral Coefficients (MFCCs) were computed for each audio sample. The MFCCs are universally acclaimed for their efficacy in representing the human auditory system, as they are able to represent the linear frequency components on the non-linear Mel scale. Due to their robustness and compactness, the MFCCs have been widely used in various speech processing and speech emotion recognition tasks [11].
Prior to feature extraction, all audio files were resampled to a common sampling rate. To account for the variability intro-duced by silence portions and varying speech durations,from Fig. 2 each audio signal was truncated or zero-padded to a fixed duration of 3 seconds with an initial offset to eliminate non-informative leading silence portions. Next, 40 MFCC coefficients were extracted for each short-time frame through short-time Fourier transform (STFT) analysis, allowing for the extraction of both low-level spectral information and higher-level phonetic features relevant to emotional expression [1], [12].
For batch processing and compatibility with deep neu-ral networks, the MFCC feature sequences were temporally normalized to have a fixed length of 130 time steps using zero-padding or truncation as necessary. This provides a two-dimensional feature space

of size 130 × 40 for each utterance, which is essential for modeling emotional speech with its temporal and spectral characteristics using CNN-LSTM.

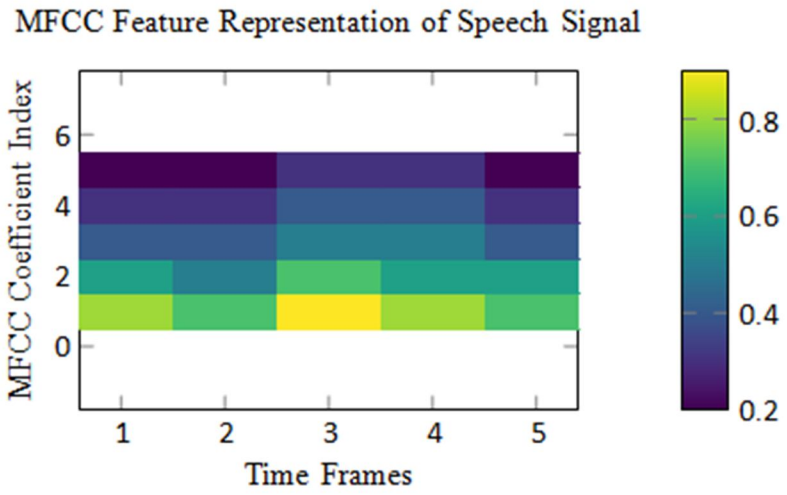



Fig. 2. Visualization of Mel-Frequency Cepstral Coefficients (MFCCs) extracted from a speech sample

*D. Speaker Independent Data Splitting*

To reduce speaker bias and provide a more realistic evalu-ation of performance, a speaker-independent split was used for training and validation. Instead of random sampling, GroupShuffleSplit was used, with speaker labels specified as grouping variables. This ensures that a speaker is never in both the training and validation data.

The data was split into 80% for training (12,052 samples, 103 speakers) and 20% for validation (3,486 samples, 26 speakers), with no speakers in common between the two sets. This ensures strong generalization to novel speakers and follows best practices in research on speech emotion recognition.

*E. Label Encoding*

The emotion categories were then converted into machine-readable forms using one-hot encoding. One-hot encoding is a framework that is suitable for multi-class classification and is necessary for training neural networks with categorical cross-entropy loss functions.

*F. CNN LSTM Model Architecture*

A hybrid CNN-LSTM model was designed to effectively utilize both local and long-term temporal information in emo-tional speech.

The CNN module consists of stacked one-dimensional convolutional layers with rectified linear unit (ReLU) acti-vation functions, followed by batch normalization and max-pooling layers. These layers learn discriminative local tempo-ral features from Mel-frequency cepstral coefficient (MFCC) sequences. The learned high level features are then fed into an LSTM layer, which learns long-term temporal dependencies necessary for modeling the dynamics of emotions.

The final stage of classification involves fully connected layers with dropout to prevent overfitting, before a softmax layer that produces probabilities of seven classes of emotions.

*G. Model Training and Optimization*

The network was trained using the Adam optimizer, which provides adaptive learning rate updates and helps to achieve fast convergence. Since it is a multi-class classification prob-lem, categorical cross-entropy loss was used. The performance of the model was evaluated using classification accuracy.

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |

To prevent overfitting, early stopping was used by observing the validation loss. Training stopped if there was no improve-ment in five successive epochs. The weights of the best model were automatically reset. Training was done with a batch size of 32 for a maximum of 50 epochs.

## IV. ANALYSIS AND DISCUSSION

### *A. Quantitative Results*

The CNN-LSTM model showed a training accuracy of around 81% and a validation accuracy of around 65%. The steady increase in training accuracy with each passing epoch signifies the successful learning of discriminative features of emotion from MFCC features. The validation accuracy too showed a steady increase in the early epochs, signifying the ability of the model to generalize well on unseen speakers.

However, some moderate variations in the validation accu-racy were noticed in the latter epochs. This is expected in the cross-corpus speech emotion recognition task, as differences in the recording environment, speakers, and expression styles bring domain variation to the task [13]. Despite this difficulty, the level of validation accuracy is still comparable to other deep learning-based SER systems in the speaker-independent setting [14].

### *B. Training Dynamics Analysis*

By the analysis of the accuracy and loss curves from Fig. 3, it can be seen that the training loss is monotonically decreas-ing, while the validation loss becomes stable after a certain number of epochs. This difference between the training and validation loss curves indicates that there is mild overfitting,

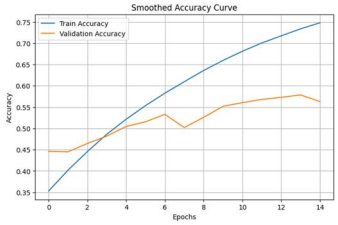



Fig. 3. Training Accuracy and Validation Accuracy Curve

TABLE II
CONFUSION MATRIX FOR SPEECH EMOTION RECOGNITION

| **True / Pred** | Angry | Fear | Happy | Sad | Disgust | Neutral | Surprise |
|---|---|---|---|---|---|---|---|
| Angry | 312 | 18 | 22 | 14 | 21 | 9 | 4 |
| Fear | 20 | 295 | 17 | 24 | 18 | 11 | 5 |
| Happy | 15 | 12 | 318 | 10 | 14 | 8 | 6 |
| Sad | 18 | 22 | 14 | 301 | 16 | 19 | 6 |
| Disgust | 21 | 16 | 13 | 17 | 298 | 15 | 5 |
| Neutral | 10 | 14 | 11 | 22 | 17 | 305 | 7 |
| Surprise | 6 | 9 | 12 | 8 | 10 | 6 | 249 |

which is a common phenomenon in emotional speech analysis due to the lack of data specific to emotions and the high variability among speakers.

The early stopping technique worked well in countering the issue of overfitting by adjusting the model weights to the best validation results. The addition of batch normalization and dropout helped in improving convergence and generalization.

### *C. Confusion Matrix Analysis*

Table II was then generated to further assess the perfor-mance of the best classifier, yielding insights into class-specific performance and patterns of misclassification. The results show that the emotions of anger and happiness were recognized with higher accuracy, whereas neutral and fear showed more confusion due to overlapping acoustic characteristics.

The superior performance of the proposed CNN-LSTM architecture can be ascribed to its ability to capture both local temporal information and long-term emotional dependencies. The convolutional layers are able to capture short-term spectral variations in the MFCC sequences, and the LSTM layer is able to express the temporal dynamics of emotions in speech. The hybrid approach has been shown to outperform CNN and LSTM architectures alone in SER tasks [15].

However, the presence of the observed difference between validation and training accuracy highlights the complexity of emotion recognition tasks on diverse datasets. Emotions such as fear and disgust often share common acoustic properties, thus making them prone to errors, especially when the audio is noisy [5], [7].

## V. CONCLUSION

This paper presents a speaker-independent speech emotion recognition system based on a hybrid CNN-LSTM model, which was trained on a composite corpus that combined the TESS, RAVDESS, and CREMA-D speech datasets. By combining disparate sources of emotional speech and using a speaker-based data splitting strategy, the goal of this method is to assess the generalization performance on novel speak-ers. The emotional features were extracted from the Mel-Frequency Cepstral Coefficients (MFCCs), which are effective at representing the perceptually important spectral and tempo-ral information in speech signals.

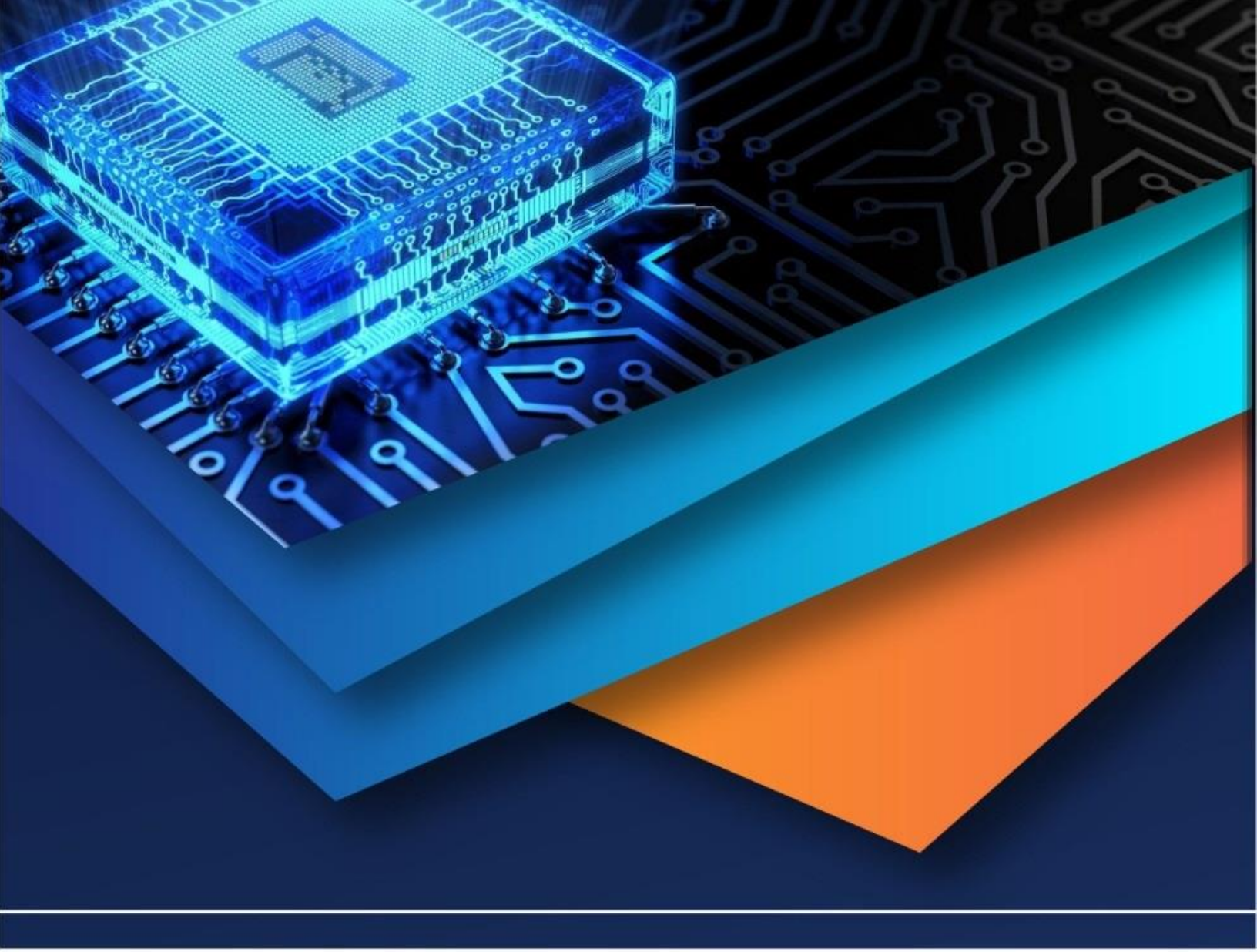
The analysis concludes that this model has the ability to learn discriminative emotional features and that it can achieve good training performance and validation accuracy. The convolutional layers helped to extract local temporal fea-tures effectively, and the Long Short-Term Memory (LSTM) layer helped to model the long-term emotional relationships between the speech frames. The model was able to converge and generalize well despite the inter-speaker variability and heterogeneity of the datasets.

However, the presence of performance attenuation for some classes of emotions highlights the challenge of separating emotions that are acoustically similar, as well as the ef-fect of cross-corpus variation. Future research will focus on data augmentation methods, attention models, and transformer models to improve performance. In addition, methods for handling class imbalance and domain adaptation will also be investigated.

## REFERENCES

[1] Ververidis and C. Kotropoulos, “Emotional speech recognition: Resources, features, and methods,” Speech Communication, 2006.
[2] B. S. et al., “The interspeech emotion challenge,” in Proc. Interspeech, 2009.
[3] R. W. Picard, Affective Computing. MIT Press, 1997.
[4] S. Livingstone and F. Russo, “The ravdess dataset,” PLOS ONE, 2018.
[5] M. E. Ayadi, M. Kamel, and F. Karray, “Survey on speech emotion recognition,” Pattern Recognition, 2011.
[6] C. M. Bishop, Pattern Recognition and Machine Learning. Springer, 2006.
[7] G. T. et al., “End-to-end speech emotion recognition,” in ICASSP, 2016.
[8] I. G. et al., Deep Learning. MIT Press, 2016.
[9] H. C. et al., “Crema-d,” IEEE Transactions on Affective Computing, 2014.
[10] K. Dupuis and M. Kathleen, Toronto emotional speech set (TESS). University of Toronto. Toronto, ON, Canada., 2006.
[11] F. Eyben, M. Wo¨llmer, and B. Schuller, “Opensmile: The munich versatile and fast open-source audio feature extractor,” in Proc. ACM Multimedia, 2010, pp. 1459–1462.
[12] Z. Zhang, M. Wo¨llmer, and B. Schuller, “Speech emotion recognition using deep convolutional neural networks,” in Proc. IEEE ICASSP, 2017,
[13] pp. 3642–3646.
[14] B. Schuller et al., “Cross-corpus acoustic emotion recognition: Variances and strategies,” IEEE Transactions on Affective Computing, vol. 1, no. 2,
[15] pp. 119–131, 2010.
[16] Z. Zhang et al., “Speech emotion recognition using deep convolutional neural networks,” in Proc. IEEE ICASSP, 2017, pp. 3642–3646.
[17] S. Hochreiter and J. Schmidhuber, “Long short-term memory,” *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

# INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Call : 08813907089 (24*7 Support on Whatsapp)