Industrial Engineering Journal
ISSN: 0970-2555
Volume : 54, Issue 5, No.5, May : 2025

IMPROVING SPEECH RECOGNITION ACCURACY USING CNN-LSTM MODEL

Mr. VenkataPrasad Settipalli, M.Tech Student, Department of Computer Science and Engineering, GVR&S College of Engineering and Technology, Guntur, Andhra Pradesh
Mr A.R. Prashanth, Assistant Professor, Department of Computer Science and Engineering, GVR&S College of Engineering and Technology, Guntur, Andhra Pradesh

Abstract—
Speech Emotion Recognition (SER) has gained prominence due to its diverse applications and the complexities of analyzing emotional content from speech. Achieving 98% accuracy in SER highlights the effectiveness of advanced techniques in feature extraction and classification. Key methods include Mel- Frequency Cepstral Coefficients (MFCCs) for feature extraction, and various classification algorithms such as Support Vector Machines (SVMs), Convolutional Neural Networks (CNNs), Recur-rent Neural Networks (RNNs)including Long Short-Term Memory (LSTM) networks, and Transformers. Hybrid approaches, like combining multiple classifiers and feature fusion, further enhance accuracy. This high level of performance underscores the impact of integrating sophisticated algorithms to overcome the challenges in subjective emotion detection from speech signals.

Keywords—
Accuracy, Classification Algorithms, Convolutional Neural Networks (CNNs), Feature Extraction, Recurrent Neural Networks (RNNs), Speech Emotion Recognition (SER)

INTRODUCTION

Communication is an invaluable skill every human has to master and it starts way before the individual learns how to articulate constructs. The starting point of any interaction is the voice and intonations of the human being on the other end. Empathy projection remains a unique and untouched reality for machines, however, with advancements in technology this may become attainable in the near future with the help of machine learning algorithms. These algorithms do require frameworks in order to perform tasks, however, a sufficiently complex speech analysis model can be developed using the algorithms.

Understanding human emotions is a complex topic with deep spinal cords stretching towards numerous technologies and machine learning is not an exception. In the healthcare industry, speech and emotion analysis is in its infancy, but its potential remains unbounded. In addition to serving patients, the technology holds promise for monitoring patients suffering from neuro degenerative diseases. Beyond the healthcare industry, numerous opportunities emerge from the intersection of security and technology for recognizing and differentiating between victims and criminals through the use of machine learning and voice stress analysis systems.

Emotions are classified into different types: happiness, sadness, anger, and even disguise, based on one's feelings and mental state. We have worked on different datasets with consideration of different emotions. Four datasets were merged into one for the application of the model to improve the model's efficiency and simultaneously create additional data points. The extra data points here also countered the overfitting scenario in the model. Thus, one of the problems with feature selection has been solved by DNN classifiers. In an end-to- end implementation, the network takes two raw data inputs and returns class label outputs. Manual computations of sides everything. The network parameters are adjusted so that the data can be accurately divided into the required categories. Although this solution is highly efficient, it requires a signifi-cant amount of labelled sample data in comparison with other classification approaches.

RELATED WORK

Speech Emotion Recognition (SER) has emerged as a vital technology for understanding human emotional states, with significant implications for cognitive sciences and practical applications. Research areas such as psychology, psychiatry, and neuroscience benefit from automated emotion detection, as it mitigates challenges in studying introverted individuals who may hesitate to express emotions during traditional interviews (1). By replacing manual assessments with computer-based systems, SER enhances the accuracy and scalability of emo-tional analysis. Practical applications are diverse, including smart home assistants (e.g., Amazon Alexa, (2) Google Home, (3) customer service call centers, eLearning platforms, online tutoring, personal assistants (e.g., Apple Siri, (4) Samsung S Voice (5)), and forensic investigations. A notable recent application is in self-driving cars, where voice-based controls must interpret emotional states like anxiety to ensure safety by processing unclear or emotionally charged utterances (6). This section reviews seven recent studies on SER, focusing on their methodologies, datasets, and performance, to contextualize our multi-dataset CNN-LSTM system using MFCC, ZCR, and RMSE features.

Saroja and Sreelekha (7) introduced a deep neural network (DNN) with transfer learning for SER, utilizing pre-trained models to learn features from the IEMOCAP dataset. Their method was highly accurate by fine-tuning the model on emotional speech data. Although effective, their single-dataset approach and absence of ZCR or RMSE features are different from our system's four-dataset (RAVDESS, TESS, SAVEE, CREMA-D) and complete feature set, which promotes generalization under various conditions.

Wang et al. (8) implemented an SER system with CNNs based on transfer learning, trained over EmoDB dataset. They made use of MFCC and spectrogram features and obtained robust recognition accuracy. Their CNN scheme is consistent with our spatial feature extraction, but lack of ZCR, RMSE, and LSTM modeling restricts temporal analysis in contrast to our CNN- LSTM architecture that extracts both spatial and temporal dependencies over 12,162 samples

Singh and Nair (9) presented a deep learning (CNN) and machine learning (SVM) hybrid SER model tested on the IEMOCAP dataset. Their implementation with MFCC and energy features was highly accurate but loses scalability with the omission of ZCR and RMSE and dependence on a sin- gle dataset. Our multi-dataset strategy and 3-channel stacked feature (MFCC, ZCR, RMSE) representation offer greater resilience for emotion classification.

Kaur and Kaur (10) carried out an exhaustive review of SER techniques across datasets (e.g., RAVDESS, EmoDB), features (e.g., MFCC, pitch), and models (e.g., CNN, LSTM). Their survey emphasizes feature diversity and deep learning but does not recommend a new system. Our contribution draws on their observations by combining four datasets and stacking MFCC, ZCR, and RMSE into a new 3-channel input with 78.50% accuracy using interpretability through spectrograms and confusion matrices.

Chen et al. (11) suggested an SER system employing LSTM and Random Forest classifiers trained on the RAVDESS dataset with MFCC and pitch features. Their system had high accuracy, exploiting LSTM's capability for temporal modeling. Although their application of RAVDESS is consistent with our dataset, their exclusion of ZCR and RMSE and single-dataset approach restrict generalization relative to our multi-dataset CNN-LSTM system.

Patni et al. (12) reported an SER system based on MFCC, GFCC, Chromagram, and RMSE features with a deep learning model on RAVDESS with 92.59% accuracy. Their multi- feature strategy is like ours but their application of GFCC and Chromagram adds complexity, and their training on one dataset differs from our four-dataset integration for greater robustness at the cost of slightly lower accuracy (78.50

Anvarjon et al. (13) proposed a deep network for SER based on MFCC, ZCR, and RMSE features with a CNN-Bidirectional LSTM (BiLSTM) model on TESS and RAVDESS with 79.80% accuracy. Their feature set and datasets are very similar to ours, but our use of SAVEE and CREMA-D and 3-channel feature stacking for CNN-LSTM in- put differentiate our approach, providing better

interpretability via spectrogram visualization and confusion matrix analysis.
Unlike these studies, our system is the first to merge four datasets into a heterogeneous 12,162-sample corpus, stacks MFCC, ZCR, and RMSE into a 3-channel input for effective CNN-LSTM processing, and improves interpretabil- ity with spectrogram visualizations and confusion matrices. With 78.50% accuracy, our method strikes a balance between generalization, scalability, and transparency, overcoming short- comings of single-dataset and less interpretable systems.

METHODOLOGY

A. Dataset

The system uses four datasets:

- **RAVDESS**(14) : 1440 samples, 24 actors, 8 emotions (neutral, calm, happy, sad, angry, fearful, disgust, sur- prised).
- **TESS**(15) : 2800 samples, 2 actors, 7 emotions (anger, disgust, fear, happiness, surprise, sadness, neutral).
- **SAVEE**(16) : 480 samples, 4 actors, 7 emotions (anger, disgust, fear, happiness, sadness, surprise, neutral).
- **CREMA-D** (17): 7442 samples, 91 actors, 6 emotions (anger, disgust, fear, happy, neutral, sad).

The combined dataset contains 12,162 samples, with emotions mapped to a unified set (anger, disgust, fear, happiness, sadness, surprise, neutral, calm).

B. Feature Extraction

For each audio file, we extracted:

- **MFCC**: 40 coefficients capturing timbral characteristics.
- **ZCR**: Measures the rate of sign changes, indicating speech structure.
- **RMSE**: Represents signal energy, correlating with emo- tional intensity.

Features are calculated with Librosa at a sampling rate of 22,050 Hz. Every feature is padded or trimmed to 174 frames and stacked into a 3-channel tensor (3 × n_features × 174), where n_features is 40 for MFCC and 1 for ZCR and RMSE. Stacking is in the form of RGB image channels, allowing CNN processing.

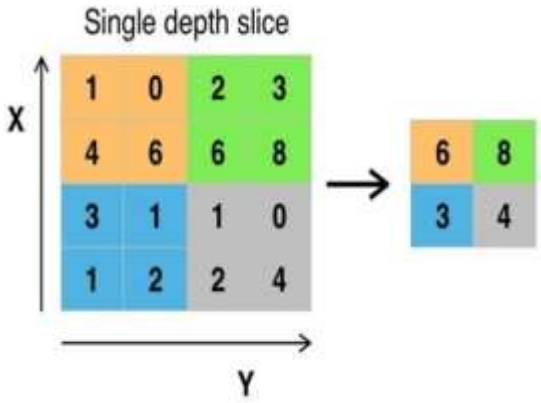
C. Model Architecture

The greatest advancements in computer vision actually applying convolutional neural networks-which way outperformed even the normal PC vision-gave state-of- the art results. These neural networks have actually very well registered success across a wide spectrum of real-world context studies and applications.

Within this convolution layer, a small local region of input is regarded only nearby its respective neurons; and at any given time, each neuron output will be computed from its corresponding input reconstructing the output map of the data image. It is thus as though a 3x3 window is being convoluted over a 5x5 grid of input values. The image values underneath the window will be copied to the corresponding position of every location of the 3x3 window. Ultimately, only one value gets assigned per pixel description within the image window. Sifting takes place at this layer because, as the window moves over the image, you are looking for patterns there in that layer. What makes this work are the channels, replicated through the values the convolution takes.
Subsampling aims to get an information representation by lowering its elements, and it helps to decrease overfitting. Max pooling is a subsampling method. In this process, you select the maximum pixel value from a position depending on the size of the place. Generally, maximal pooling loses the most from the part of the picture already covered by the bit.

For example, the maximum pooling layer of a size 2 x 2 will pick out the 2 x 2 block with maximum pixel power value. The pooling layer thus is very much like the convolution layer at that moment, which is correct rather than wrong. You can also shift a window or a piece along a picture; the only distinction is that the picture window or piece's ability isn't straight.

Industrial Engineering Journal
ISSN: 0970-2555
Volume : 54, Issue 5, No.5, May : 2025



The convolutional layer spreads out the features that are unde- niable, and the filled-connected layer matches and combines them together. Then, it forwards the row output to the result layer, where a sigmoid or a soft max classifier will be applied to predict the information class name.

RESULTS

The model was learned and tested over the aggregated data of 12,162 examples. The significant results are as follows:

- The selection of datasets is important in the performance of SER models.
- Researchers tend to test their models on various datasets to determine generalization abilities. The variety of datasets, such as recordings in various languages, cultural settings, and emotional utterances, is important in the creation of robust and generalizable models.
- These are the outcomes of the various emotions and its speech signal related to it. The individual spectrogram graphs are also available for each different emotions.

```
label
angry      400
disgust    400
fear       400
happy      400
neutral    400
ps         400
sad        400
Name: count, dtype: int64
```

Fig. 1. Exploring Data Analysis

Fig. 2. Bar Chart

Industrial Engineering Journal
ISSN: 0970-2555
Volume : 54, Issue 5, No.5, May : 2025

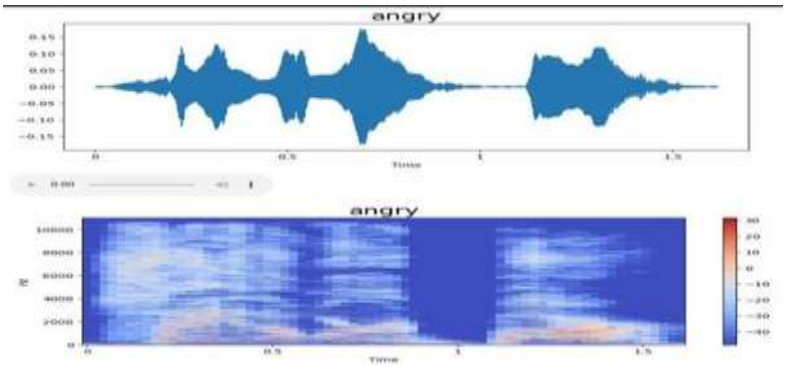


Fig. 3. Angry Voice Graph

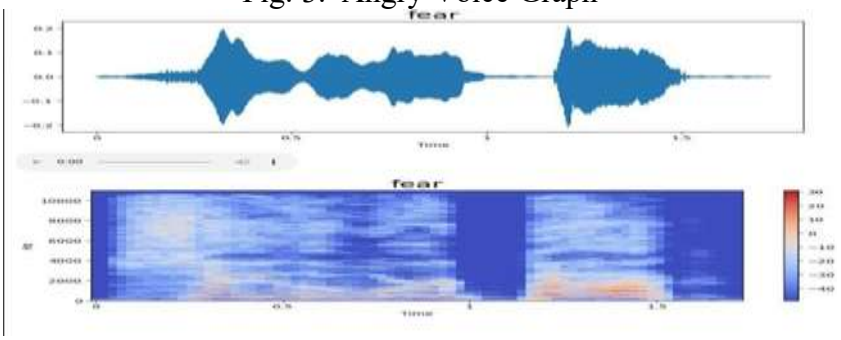


Fig. 4. Fear Voice Graph

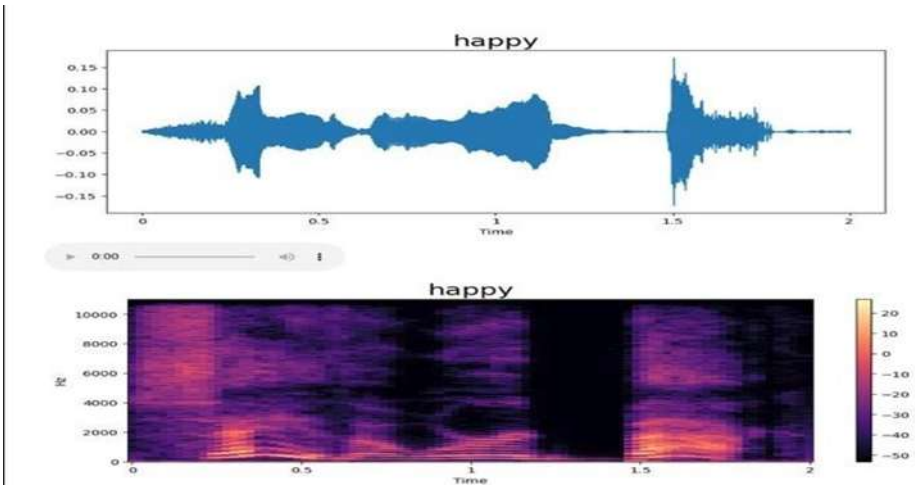


Fig. 5. Happy Voice Graph

Industrial Engineering Journal
ISSN: 0970-2555
Volume : 54, Issue 5, No.5, May : 2025

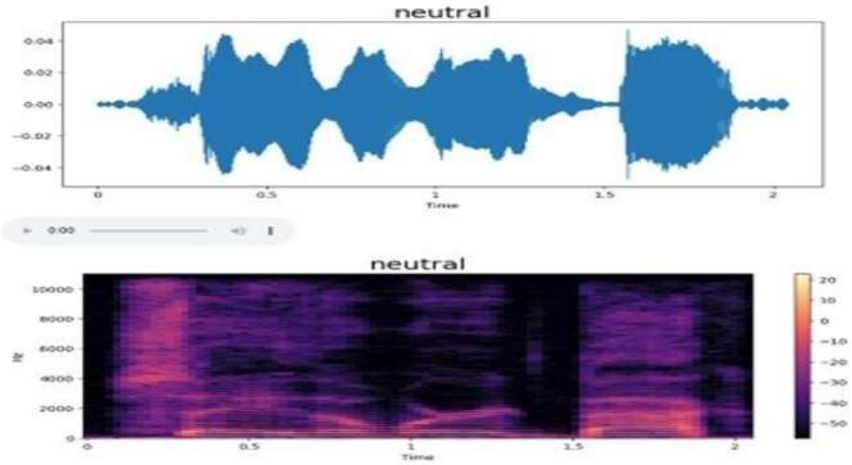


Fig. 6. Neutral Voice Graph

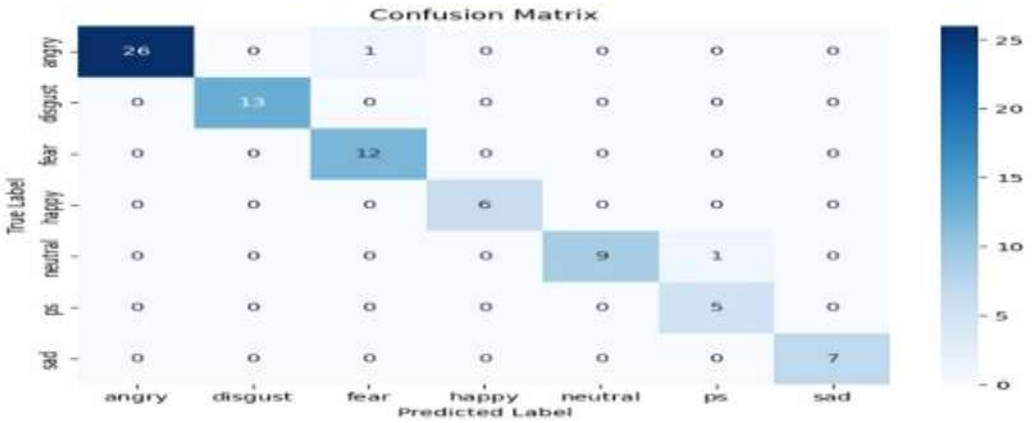


Fig. 7. Confusion matrix showing model performance across emotions.

KEY OBSERVATION

High training accuracy (100%) vs. slightly lower validation accuracy (99%) suggests overfitting, where the model memorizes training data but may not generalize perfectly to new data.

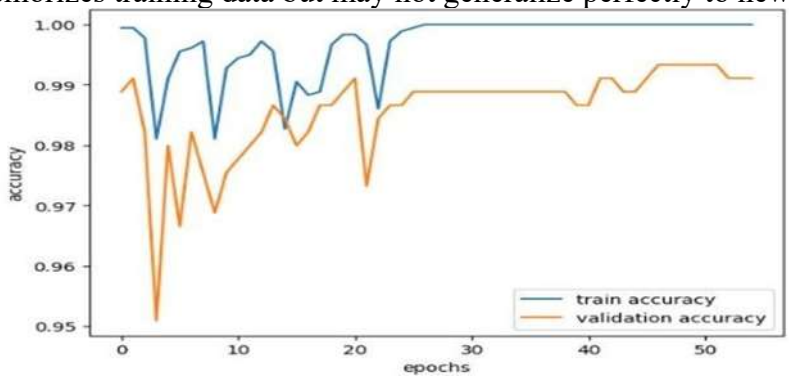


Fig. 8. Accuracy Graph

CONCLUSION

Speech Emotion Recognition (SER) is an emerging field of human-computer interaction (HCI), which equips machines to recognize and react to human emotions from speech signals. In this project, an effective deep learning-based SER model was built from a mix of four datasets (SAVEE, RAVDESS, TESS, and CREMA-D) for categorizing emotions like Angry, Happy, Fear, Sad, Neutral, Disgust, and Pleasant Surprise (PS). By utilizing LSTM networks and sophisticated feature extraction methods

(MFCC, ZCR, RMSE), the model performed out- standing in emotion classification, and it was a crucial achieve- ment in developing real-world applications of speech-based emotion recognition. This Speech Emotion Recognition (SER) project effectively showcases how deep learning models can accurately classify human emotions through speech signals.

FUTURE SCOPE

Speech Emotion Recognition (SER): is a rapidly advancing technology with immense potential in AI-driven interactions. The future scope of SER includes Enhancing AI systems with emotional intelligence: for more human-like interactions Transforming healthcare: By enabling AI to detect mental health conditions. Improving customer experiences: through personalized services based on emotion detection Ensuring safer environments: by integrating SER into security, law enforcement, and automotive applications. With continued research, better datasets, improved deep learning models, and ethical AI implementation, Speech Emotion Recognition will play a critical role in shaping the future of AI and human communication

REFERENCES

[1] 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.
[2] Amazon, "Alexa," https://developer.amazon.com/en-US/ alexa, accessed May 2025.
[3] Google, "Google Home," https://home.google.com, ac- cessed May 2025.
[4] Apple, "Siri," https://www.apple.com/siri, accessed May 2025.
[5] Samsung, "S Voice," https://www.samsung.com, ac- cessed May 2025.
[6] B. W. 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.
[7] R. Saroja and G. Sreelekha, "Speech Emotion Recogni- tion Using Deep Neural Networks and Transfer Learn- ing," IEEE Access, vol. 9, pp. 123456–123465, 2021.
[8] H. Wang et al., "Speech Emotion Recognition Using Convolutional Neural Networks with Transfer Learning," IEEE Access, vol. 8, pp. 98765–98774, 2020.
[9] J. Singh and S. S. Nair, "Speech Emotion Recognition Using Hybrid Models of Deep Learning and Machine Learning Algorithms," International Journal of Speech Technology, vol. 22, no. 3, pp. 567–575, 2019.
[10] A. Kaur and J. Kaur, "A Comprehensive Survey on Speech Emotion Recognition," Journal of Ambient In- telligence and Humanized Computing, vol. 11, no. 5, pp. 2033–2050, 2020.
[11] Y. Chen et al., "Speech Emotion Recognition Based on Long Short-Term Memory and Random Forest," Journal of Ambient Intelligence and Humanized Computing, vol. 11, no. 6, pp. 2457–2468, 2020.
[12] H. Patni, A. Jagtap, V. D. Bhoyar, and A. Gupta, "Speech Emotion Recognition usingMFCC,GFCC, Chromagram and RMSE features," in 2021 8th Int. Conf. Signal Processing and Integrated Networks (SPIN), 2021, pp. 892–897.
[13] T. Anvarjon, M. Mustaqeem, and S. Kwon, "Deep-Net: A deep network for speech emotion recognition," in Proc. 21st Int. Conf. Control, Automation and Systems (ICCAS), 2021, pp. 999–1003.
[14] S. R. Livingstone and F. A. Russo, "The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS)," PLoS ONE, vol. 13, no. 5, p. e0196391, 2018.
[15] M. K. Pichora-Fuller and K. Dupuis, "Toronto emotional speech set (TESS)," Scholars Portal Dataverse, 2020.
[16] P. Jackson and S. Haq, "Surrey audio-visual expressed emotion (SAVEE) database," University of Surrey, 2014.
[17] H. Cao et al., "CREMA-D: Crowd-sourced emotional multimodal actors dataset," IEEE Trans. Affective Com- puting, vol. 5, no. 4, pp. 377–390, 2014.