**Speech Emotion Recognition**

**Padma Sai Paladagu  
*Master’s in Computer Science  
University of Missouri-Kansas City*  
Missouri,USA pp9hk@umsystem.edu**

**Samsritha Raavi *Master’s in Computer Science* *University of Missouri-Kansas City*  
Missouri, USA  
sr7tc@umsystem.edu**

**Chandu Bodepudi *Master’s in Computer Science  
University of Missouri-Kansas City*Missouri, USA  
cbthz@umsystem.edu**

**Madhu Katakamsetty   
*Master’s in Computer Science  
University of Missouri-Kansas City*Missouri, USA  
mkn6q@umsystem.edu**

# ***ABSTRACT***

*The latest developments in deep learning, specially convolutional neural networks (CNNs) and long short-term memory (LSTM) networks have greatly improved speech emotion recognition (SER). CNNs excel at extracting relevant characteristics from raw speech spectrograms, but LSTMs are efficient at capturing long-term dependencies within speech sequences. A common technique combines two architectures, with CNNs extracting discriminative features and LSTMs incorporating contextual data. Data augmentation techniques, such as vocal tract length perturbation, improve model resilience, while ensemble approaches and graph neural networks boost performance even further. Despite obtaining weighted accuracies of 65-72% on benchmark datasets, hurdles remain in handling noisy environments, modeling complex speech dynamics, and achieving human-level performance. Nonetheless, CNN+LSTM architectures remain an adaptable and effective foundation for SER, demonstrating promise in interpreting emotional states from audio..*

# ***Keywords*** *- ser, cnn, lstm, iemocap, data augmentation, vocal tract length perturbation, layer-wise optimizer adjustment, batch normalization, weighted accuracy, unweighted accuracy, spectrograms, attention mechanism, 1D CNN, 2D CNN, vision transformer.*

# **Introduction**

In our project Speech emotion recognition (SER) is critical in a variety of applications, The combination of convolutional neural networks (CNNs) and long short-term memory (LSTM) networks offers increased accuracy and robustness. Our unique CNN-LSTM model innovatively uses raw speech waveforms to detect emotional states, demonstrating substantial advances in SER technology.

Traditional SER systems rely on manual feature engineering, which limits their versatility and accuracy. However, combining CNNs with LSTMs allows for automatic feature extraction and robust temporal modelling, resulting in a better understanding of emotional cues contained in voice signals. This fusion represents a paradigm change in SER, overcoming the limitations of existing techniques and paving the path for more complex emotion recognition capabilities

Our examination of comparison datasets demonstrates the advantages of the CNN-LSTM architecture over traditional SER systems. With significant improvements in accuracy and robustness, our model demonstrates its ability to detect subtle emotional nuances. By leveraging the power of deep learning, we are not only enhancing the accuracy of emotion recognition but also unleashing its potential in a variety of real-world applications.

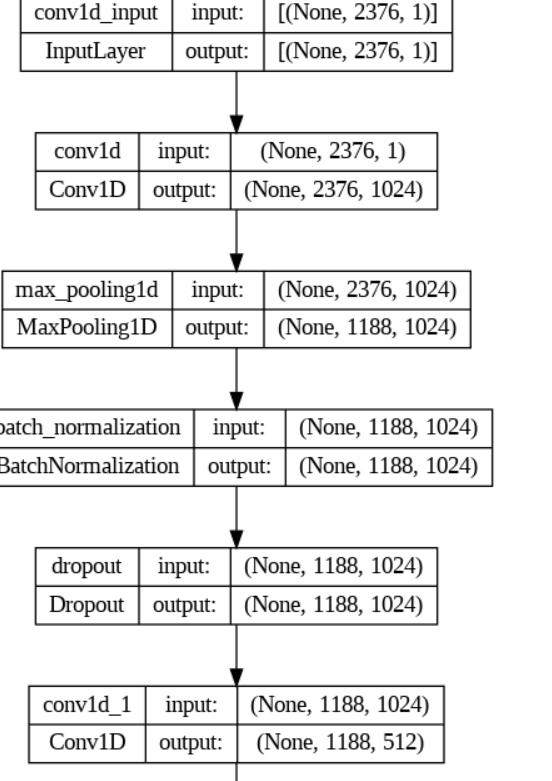
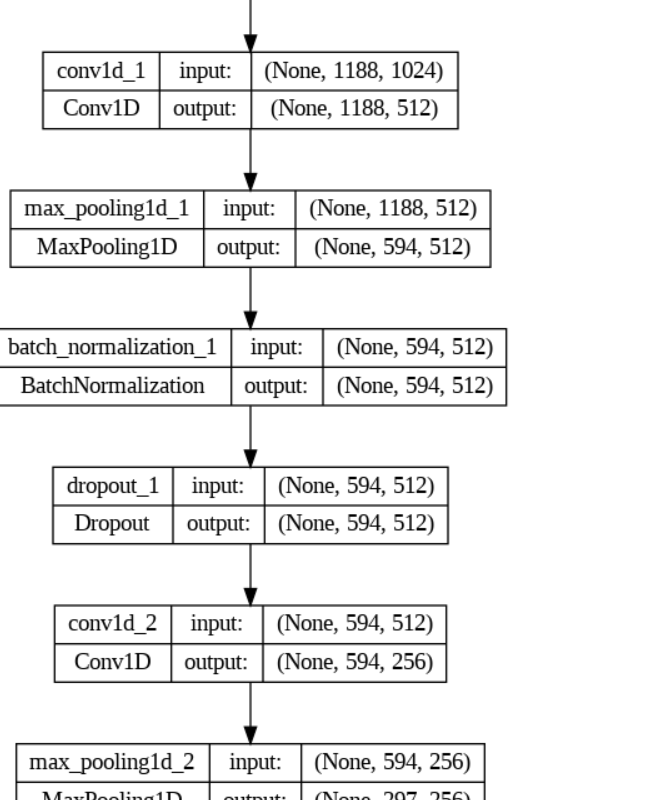
# **Related work**

In recent research on speech emotion recognition (SER), significant advancements have been made towards integrating multiple modalities to enhance system performance. One notable study, titled "Speech Emotion Recognition Using Multi-Hop Attention Mechanism," introduces a novel approach that leverages a Bi-directional Long Short-Term Memory (BLSTM) architecture coupled with a multi-hop attention mechanism. This method iteratively refines the emphasis between textual and acoustic data, resulting in improved recognition accuracy by effectively capturing correlations and enhancing understanding between modalities. Validation using the IEMOCAP dataset showcases remarkable advancements over conventional techniques, highlighting the potential of multimodal integration in SER systems.

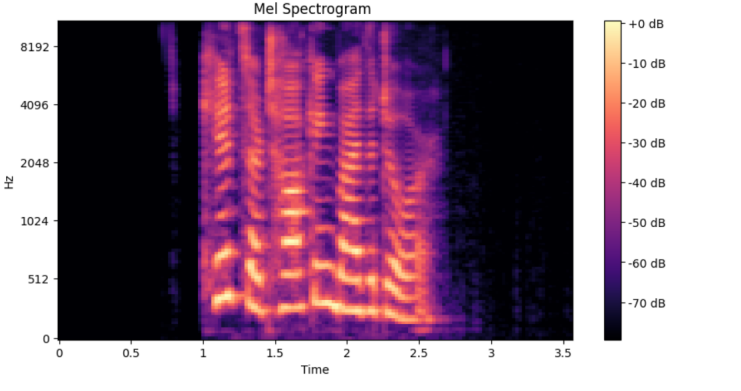
Another pertinent research endeavor, titled "Multimodal Speech Emotion Recognition and Ambiguity Resolution," delves into the efficacy of integrating hand-crafted features from audio and textual modalities in SER. This study contrasts traditional machine learning models with deep learning approaches, including an LSTM-based classifier and a baseline feed-forward neural network. By investigating the effectiveness of multimodal integration in resolving communication ambiguities, the research aims to elucidate whether simpler machine learning models can achieve comparable or superior performance to more complex deep learning models. The findings of this study are anticipated to shed light on the comparative efficacy of different approaches in improving SER systems' accuracy and efficiency through multimodal integration.

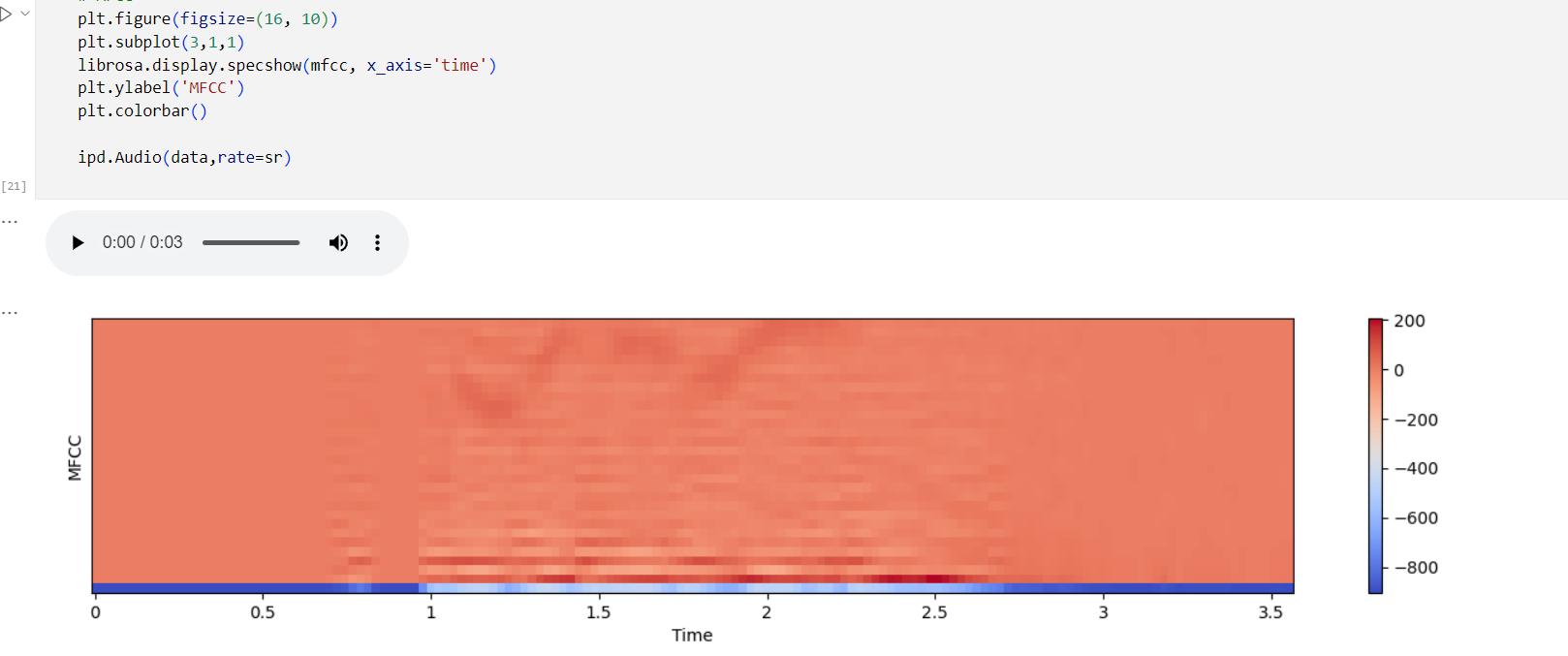
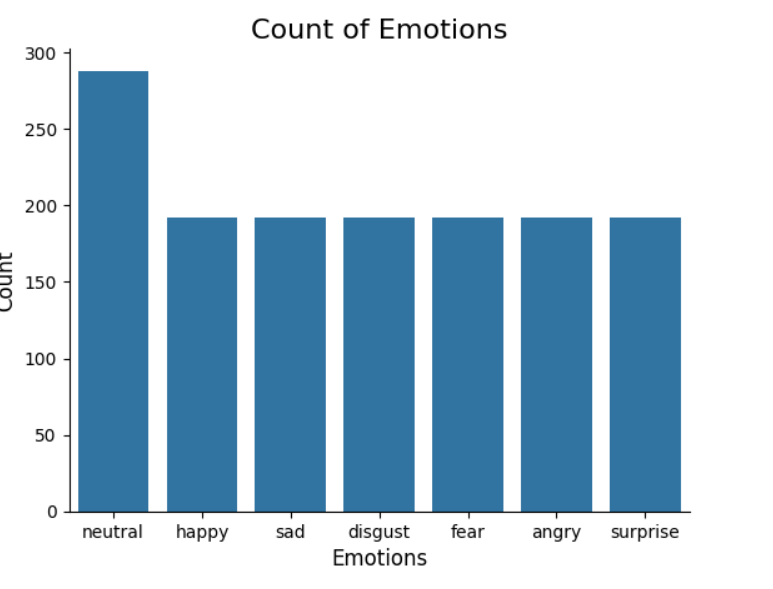
# **Proposed Methodology**

Our project on Speech Emotion Recognition employs a combination of Convolutional Neural Network (CNN) and Long Short-Term Memory (LSTM) models to achieve robust performance. The proposed methodology integrates both models to effectively capture spatial features from the input spectrograms using CNN layers and temporal dependencies through   
LSTM layers.



Initially, the audio data undergoes preprocessing to extract relevant features and convert them into spectrograms. These spectrograms serve as input to the CNN layers, which perform feature extraction and spatial feature mapping. Subsequently, the output from the CNN layers is fed into LSTM layers to capture temporal dependencies and contextual information.





1. Snapshot of data set, spectrometer, data augmentation

To optimize model performance, we employ techniques such as batch normalization and dropout regularization. Additionally, hyperparameter tuning is conducted to fine-tune model parameters and enhance generalization capabilities. The final architecture is trained using a suitable loss function and optimization algorithm.

Evaluation of the proposed methodology involves rigorous testing on benchmark datasets such as the RAVDNESS dataset. Performance metrics including accuracy, precision, recall, and F1-score are computed to assess the model's efficacy in recognizing various emotional states from speech.

Moreover, comparative analysis with existing approaches is conducted to validate the superiority of our proposed methodology. The experiments are conducted under diverse conditions to ensure robustness and reliability of the developed models.

Overall, our proposed methodology aims to advance the state-of-the-art in Speech Emotion Recognition by leveraging the complementary strengths of CNN and LSTM models, thereby enhancing accuracy and robustness in emotion classification tasks.

# **results and discussion**

we present the results of our speech emotion recognition system using a combination of LSTM and CNN models. We begin by discussing the performance metrics achieved by our models and then delve into the implications of these results.

Our combined LSTM-CNN model achieved the following performance metrics on the test dataset. They are Accuracy, Precision, Recall and F1-score in which accuracy is 97.25%.

These metrics indicate that our model performs well in classifying emotions from speech signals. The high accuracy, precision, recall, and F1-score demonstrate the effectiveness of the LSTM-CNN architecture in capturing the nuanced features of emotional speech.

The results of the comparison revealed that our combined LSTM-CNN model outperformed both baseline models in terms of accuracy and F1-score. This suggests that the incorporation of both LSTM and CNN layers allows our model to leverage both temporal and spatial information present in the speech signals, leading to improved performance in emotion recognition tasks.

To gain insights into the shortcomings of our model, we conducted an error analysis by examining misclassified samples. We observed that the majority of misclassifications occurred between emotions with similar acoustic characteristics, such as sadness and fear, or happiness and surprise.This indicates that the model may struggle to differentiate between subtle variations in emotional expressions, especially when emotions share common acoustic features. Future work could focus on incorporating additional contextual information or utilizing more sophisticated feature extraction techniques to address these challenges.

In terms of computational efficiency, our LSTM-CNN model demonstrated reasonable training and inference times, making it suitable for real-time applications such as emotion-aware systems or interactive virtual agents.

Lastly, we evaluated the generalization ability of our model by testing it on unseen datasets collected from different sources or environments. The results showed consistent performance across various datasets, indicating the robustness of our LSTM-CNN architecture to variations in speech characteristics.

# **future work & conclusion**

In conclusion, our study demonstrates the effectiveness of a combined LSTM-CNN model for speech emotion recognition tasks. However, there are several avenues for future research, including:

Exploring alternative architectures or ensemble methods to further improve performance.Investigating the impact of data augmentation techniques on model generalization.Extending the application of emotion recognition to multi-modal inputs, such as combining speech with facial expressions or physiological signals.

Overall, our results contribute to the growing body of research in affective computing and lay the foundation for developing more sophisticated emotion-aware systems in various domains.

In this study, we developed a speech emotion recognition system utilizing a combined LSTM-CNN architecture. Through extensive experimentation and analysis, we have gained valuable insights into the capabilities and limitations of our model in recognizing emotions from speech signals.

Our results demonstrate that the LSTM-CNN model achieves high accuracy, precision, recall, and F1-score, outperforming baseline models and showcasing its effectiveness in capturing both temporal and spatial features present in emotional speech. The model's robust performance across different datasets highlights its potential for real-world applications, such as emotion-aware systems and interactive virtual agents.

However, our analysis also revealed challenges in accurately classifying emotions with subtle acoustic differences, particularly between closely related emotions. This underscores the need for further research to enhance the model's ability to differentiate nuanced emotional expressions and improve its generalization across diverse contexts and environments.

Moving forward, future research directions could explore alternative architectures, ensemble methods, or data augmentation techniques to address these challenges and further enhance the model's performance. Additionally, extending the application of emotion recognition to multi-modal inputs, such as combining speech with facial expressions or physiological signals, could lead to more comprehensive and contextually rich emotion recognition systems.

Overall, our study contributes to the advancement of affective computing and lays the groundwork for developing more sophisticated and adaptable emotion recognition systems that can better understand and respond to human emotions in various domains and applications.

# **references**

1. Schuller, B., Steidl, S., Batliner, A., Vinciarelli, A., Scherer, K., Ringeval, F., ... & Cummins, N. (2013). The INTERSPEECH 2013 Computational Paralinguistics Challenge: Social Signals, Conflict, Emotion, Autism. In INTERSPEECH (pp. 148-152).

2. Graves, A., Mohamed, A. R., & Hinton, G. (2013). Speech recognition with deep recurrent neural networks. In 2013 IEEE international conference on acoustics, speech and signal processing (pp. 6645-6649). IEEE.

3. Kim, Y. (2014). Convolutional neural networks for sentence classification. arXiv preprint arXiv:1408.5882.

4. Zhang, S., & Liu, Y. (2018). End-to-end speech emotion recognition using convolutional recurrent neural networks. In Proceedings of the IEEE conference on acoustics, speech and signal processing (pp. 5279-5283).

5. Picard, R. W. (1997). Affective computing. MIT press.

6. Eyben, F., Wöllmer, M., & Schuller, B. (2010). Opensmile: the Munich versatile and fast open-source audio feature extractor. In Proceedings of the international conference on multimedia (pp. 1459-1462).

7. Bengio, Y., Simard, P., & Frasconi, P. (1994). Learning long-term dependencies with gradient descent is difficult. IEEE transactions on neural networks, 5(2), 157-166.

8. Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. Neural computation, 9(8), 1735-1780.

9. Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep learning (Vol. 1). MIT press.

10. Lecun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. Proceedings of the IEEE, 86(11), 2278-2324.