International Journal of Information Technology, Research and Applications (IJITRA)

Samyuktha S, Sarwath Unnisa, (2025). Emotional Speech Recognition Using CNN Model, 4(1), 30-38.

ISSN: 2583-5343

DOI:10.59461/ijitra.v4i1.164

The online version of this article can be found at:
https://www.ijitra.com/index.php/ijitra/issue/archive

Published by:
PRISMA Publications

International Journal of Information Technology, Research and Applications (IJITRA) is a journal that publishes articles which contribute new theoretical results in all the areas of Computer Science, Communication Network and Information Technology. Research paper and articles on Big Data, Machine Learning, IOT, Blockchain, Network Security, Optical Integrated Circuits, and Artificial Intelligence are in prime position.

PRISMA
PUBLICATIONS

https://www.prismapublications.com/

Emotional Speech Recognition Using CNN Model

Samyuktha S[1], Sarwath Unnisa[2]

[1,2]Department of Computer Science, Mount Carmel College, Autonomous, Bangalore, Karnataka, India

Article Info

Article history:

Received February 15, 2025
Revised March 20, 2025
Accepted March 25, 2025

Keywords:

Speech Emotion Recognition
Mel-Spectrogram
MFCCs
Convolutional Neural Networks
Deep Learning
Data Augmentation
Affective Computing

ABSTRACT

Speech Emotion Recognition (SER) is a new area of artificial intelligence that deals with recognizing human emotions from speech signals. Emotions are an important aspect of communication, affecting social interactions and decision-making processes. This paper introduces a complete SER system that uses state-of-the-art deep learning methods to recognize emotions like Happy, Sad, Angry, Neutral, Surprise, Calm, Fear, and Disgust. The suggested model uses Mel-Spectrograms, MFCCs, and Chroma features for efficient feature extraction. Convolutional layers are utilized to capture complex patterns in audio data, whereas dropout layers are included to avoid overfitting and promote model generalization. Data augmentation strategies, such as pitch shifting, noise injection, and time-stretching, are adopted to increase model robustness. Despite improvements in SER, issues like the differentiation of closely correlated emotions, dealing with noisy environments, and real-time performance are domains for future work. This paper advances the research area of affective computing by enhancing emotion recognition performance and widening the scope of SER applications in healthcare, virtual assistants, and customer service systems.

Corresponding Author:

Samyuktha S
Department of Computer Science
Mount Carmel College, Autonomous
Bangalore, India

Email: Samyuktha6665@gmail.com

1 Introduction

Emotional speech recognition identifies human emotions from speech signals, with applications in healthcare, particularly for conditions such as Parkinson's disease (PD) and depression. PD is a movement disorder caused by the loss of dopamine-producing neurons, leading to motor symptoms such as tremors, bradykinesia, and rigidity, as well as non-motor symptoms like depression and sleep disturbances.

Speech-emotion recognition relies on tone, pitch, and speech patterns, enabling machines to understand human emotions. Traditional methods use handcrafted features such as MFCCs, chroma features, and Mel-spectrograms. However, Convolutional Neural Networks (CNNs) have shown superior performance in analyzing raw audio and spectrograms due to their ability to effectively capture emotional variations.

The fields of thinking and psycholinguistics on condition that interesting results about how prosodic cues, essential rate of recurrence, and the concentration of the voice can show unpredictability levels across different speakers. Short-term spectral features and sound quality can make known emotional pointers [1]. Identifying emotional conditions in speech signals is a challenging area for several reasons. First, the issue of all speech emotional methods is selecting the best features, which are controlling enough to distinguish between different emotions. The existence of various languages, accents, sentences, speaking style, and speakers also adds another difficulty because these characteristics in a straight-line change most of the take-out features, including pitch, energy [2]. Furthermore, it is possible to have more than one specific emotion at the same time in the same speech signal, each emotion associates with a different part of the speech signal.[2, 13]

The significant challenges of recognizing a speaker's emotional states from the speech are driven by a variability of factors. First, which speech features are most effective in cultivating distinct emotional states is not clear. The sound variability introduced by the existence of different sentences, speakers, speaking styles, and speaking rates improves additional layers of difficulty as these factors could have a through impact on the recovered speech features. The reliance of certain emotional expressions of the speaker and the speaker's culture, dialect, and environment could also affect the SER performance. Second, there may be emotion over laps or multiple emotions perceived in the same noise, making it difficult to determine the boundaries between each distinct emotional state. Even though many research efforts in SER have explored assorted ML approaches with various recipes of speech features, the widely held view of these works did not describe the techniques or methods used in carrying out the three core steps (i.e., data pre-processing, feature extraction) of the SER task. Moreover, the challenges associated with these methods, such as the prevalent low classification-accuracy issue of Speaker-Independent SER systems, and potential solutions are either not addressed or thinly deliberated.[3, 10]

In our study, we trained a model using the Ravdess dataset to classify emotions such as happiness, sadness, anger, disgust, neutrality, fear, surprise, and calmness, achieving high accuracy in both the training and testing phases.

2 Literature Survey

The area of Speech Emotion Recognition (SER) has made tremendous progress over the last few years, owing to its prospects in healthcare, virtual assistants, and customer care systems. [16]The initial studies mainly used conventional machine learning models like Support Vector Machines (SVM), Random Forests, and k-nearest Neighbors (k-NN), which needed heavy feature engineering with methods like Mel-Frequency Cepstral Coefficients (MFCCs), Zero-crossing Rate (ZCR), and spectral features. Although these techniques did moderately well, the advent of deep learning made a sea change in SER with the possibility to automatically extract features from raw audio data. Models like Convolutional Neural Networks (CNN), Recurrent Neural Networks (RNN), and Gated Recurrent Units (GRU) have then showcased better performance in extracting subtle temporal and spectral cues in an audio signal. Also, hybrid models integrating CNN and LSTM layers have yielded encouraging results in both accuracy improvement and robustness.[18]

SVM with radial basis function (RBF) is accepted. Berlin and Vera are Mittag (VAM) are hired to evaluate MSFs. In the experimental result, the MSFs display accomplished performance in comparison with MFCC and perceptual linear prediction coefficients (PLPC). When MSFs utilize enhanced prosodic features, there is a significant improvement in the presentation of recognition. Furthermore, an overall recognition rate of 91.6% is reached for the classification.[4, 6, 15]

Some of the core concepts have played an important role in furthering SER research. Feature Extraction methodologies like MFCCs, chroma features, and Mel-spectrograms are extensively used for speech characteristics like pitch, tone, and rhythm extraction. Deep learning models like Conv1D, Bi-GRU, and Transformer Networks have become widely used for modeling sequential data and detecting patterns in emotions. Additionally, data augmentation methods like noise injection, pitch shifting, and time stretching have enhanced model generalization by increasing training data diversity. Transfer learning techniques have also been increasingly adopted, using pre-trained models to enhance SER performance under limited labeled data.[17, 19, 20]

The meting out of speech signals is accomplished based on small sections with non-overlapping portions. This method was tested on the IEMOCAP dataset and reached a recognition accuracy of 68% as soon as the deep network was mutual with a high-complexity convolutional LSTM. [5]There are 2 stages in creating the mutual deep CNN. The 2 considered architectures' hyperparameters are selected using Bayesian optimization in training. Next designing and estimating One 1D CNN and One 2D CNN architecture, the 2 CNN architectures were mutually removing these 2nd layers. Transfer learning was additional to the training to help speed up the training of the mutual CNN. The 1st 2 CNNs to be trained were the 1D and 2D CNNs. The the1D and 2D CNNs learned features were then repurposed and converted to the mutual CNN.[5, 7, 8, 9]

Despite these advances, there are significant gaps in the available literature. The main challenge lies in the lack of diversities in databases with adequate representation of different languages, accents, and cultural variations, ensuring generalization for SER models. Further, although deep learning models have excellent performance, they have computational complexity that challenges their deployment in real-time in resource-constrained environments. Another major constraint includes the inexact differentiation of strongly related emotions. For instance, anger or frustration and happiness or excitement frequently

possess similar speech features. The algorithm also generally finds it difficult under noisy environments since SER models face a number of challenges and constraints in many areas.[11, 12, 14] While it has held high theoretical promises in its practical realization, such real-world applications in healthcare systems, smart assistants, and even the emotion-aware Internet of Things, for instance, are still in relative infancy. Sealing these loopholes through better diversity of datasets, advanced noise abatement methods, and effective model architectures is indispensable for the improvement of SER studies and its everyday applicability.

3 Methodology

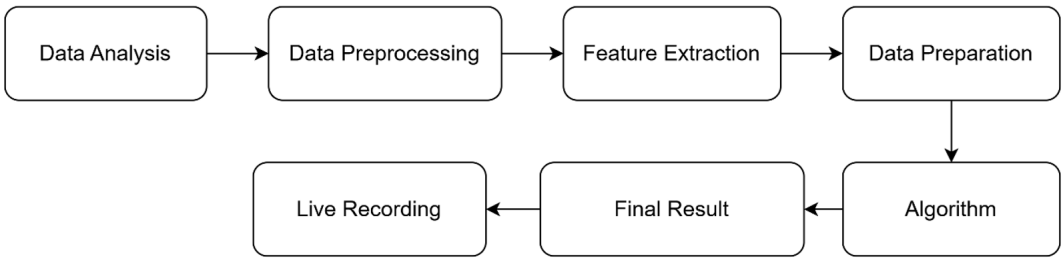

Figure 1: End-to-end Data Processing Pipeline

Figure 1 shows an end-to-end data processing pipeline that is often employed in operations such as Speech Emotion recognition (SER) and other machine learning tasks. The process starts with Data Analysis, where the gathered data is analyzed to identify its structure, identify missing values, and extract patterns. Data visualization is also often part of this step, which helps discover insights that drive the preprocessing process.

Second, in Data Preprocessing, raw data is cleaned and transformed to make it more quality-rich. It encompasses methods such as normalization, noise removal, and data augmentation (e.g., pitch shifting or time-stretching) to increase the size of the dataset and enhance model performance. After preprocessing, the data flows into the Feature Extraction stage, where important features are extracted to assist the model in identifying patterns. For audio data, these would typically consist of MFCCs, chroma features, Mel-spectrograms, and Zero-Crossing Rate (ZCR), which reflect important speech features.

Data Preparation comes next, where data for training and testing is arranged. This involves dividing the data into training, validation, and test sets, label encoding for classification, and shuffling data to minimize model bias. In real-time applications, Live Recording is also an essential step where fresh audio input is received and processed in real-time for emotion detection or other predictions.

The pre-processed data is then sent to the Algorithm stage, where the machine learning or deep learning model is used for classification. Examples of models are CNNs to classify Mel-Spectrograms, RNN/GRU/LSTM to classify sequential data, and normal models such as SVM or Random Forest for classification based on features.

Lastly, the Final Result step includes the forecasted results, e.g., emotion labels (e.g., Happy, sad, Angry, disgust, neutral, surprise, fear, calm), as well as confidence scores and performance measures like accuracy or confusion matrix. This systemized workflow guarantees a convenient and efficient process for the management of complex data-driven tasks.

4 Architecture

Speech Emotion recognition (SER) using deep learning is a multi-step process to transform and audio signal into useful emotion predictions. It begins with the Audio Signal, the initial input that is a depiction of sound in digital or analog form. Such audio signals are significant in a lot of applications such as speech recognition and music processing.

Next, Preprocessing is done to get the audio data ready for feature extraction. The audio signal is transformed into a Mel-Spectrogram, a time-frequency representation that highlights perceptually significant frequencies, making it especially suited for SER tasks. Data Augmentation operations like addition of Noise, Pitch shifting, application of Stretch, and Shift are typically done at this stage to enhance model generalization.

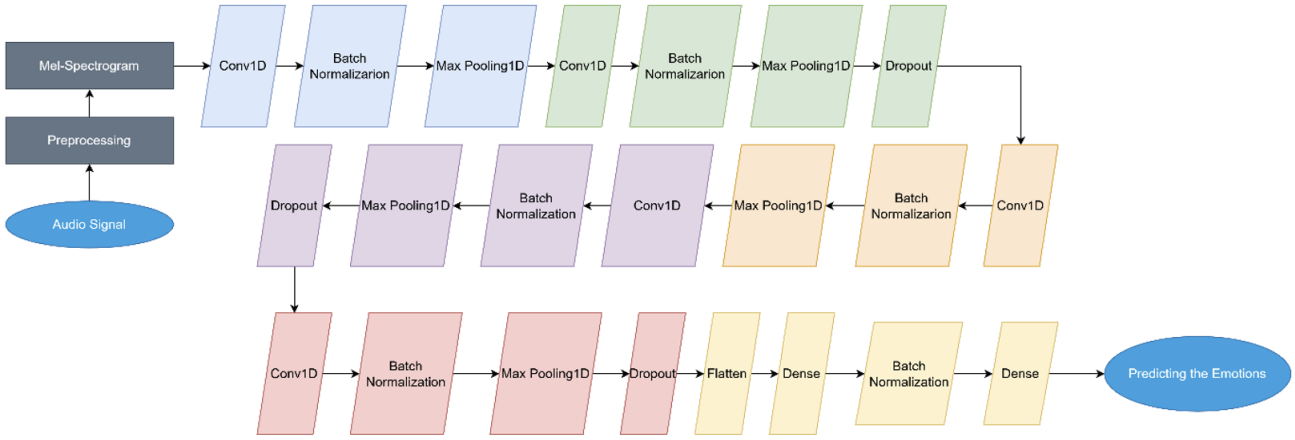

Figure 2: Speech Emotion Recognition

The third one is feature extraction through convolutional layers. The first convolutional block utilizes a Conv1D layer to extract low-level features like pitch, tone, and spectral content. This is followed by Batch Normalization to regularize learning and enhance the convergence rate. A MaxPooling1D layer is utilized for the down-sampling feature maps without losing critical information. (Fig. 2) In the second convolutional block, the same layers (Conv1D, Batch Normalization, and MaxPooling1D) are utilized to extract deeper features like intensity variations and pattern detection. To prevent overfitting, a Dropout Layer is incorporated, which randomly deactivates some neurons while training.

Deeper feature extraction is done using another Conv1D layer to identify high-level patterns, with Batch Normalization for stable learning and MaxPooling1D to continue reducing the spatial dimensions. Another Dropout layer is used to increase model robustness and avoid overfitting.

5 Results

5.1 Confusion Matrix

This confusion matrix shown in Figure 3 signifies the presentation of a classification model intended to guess emotions from speech data. Each row matches the actual labels, while each column shows the projected sticky tag. The diagonal values indicate correct guesses, while off-diagonal values represent misclassification. Higher values along the diagonal validate better model accuracy.

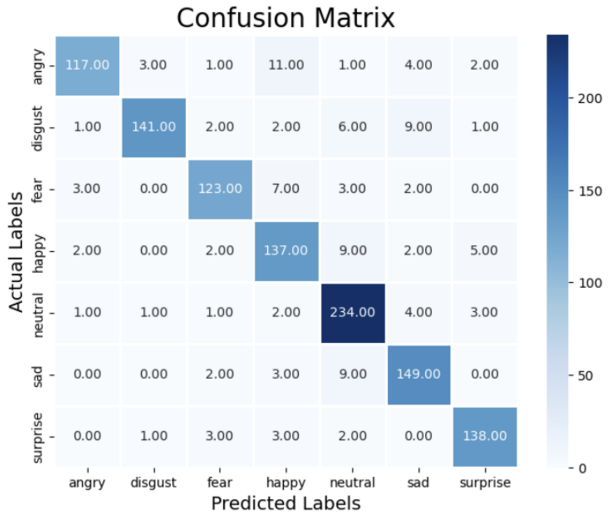

Figure 3: Confusion Matrix

The confusion matrix indicates that the model performs extremely well in classifying the neutral emotion with 234 correct classifications, the highest of all the feelings. Similarly, the model performs well in the classification of the surprise with 138 correct classifications, with hardly any confusion with other emotions. The sad emotion is also properly classified, achieving 149 correct placements.

However, there are distinguishing misclassifications, such as the model struggles with distinguishing between angry, fearful, and happy. Exactly, the model misclassified 11 'angry' trials as 'happy' and 7 'fear' trials as 'happy'. These errors indicate that these emotions may share an overlying characteristic, making them harder for the model to distinguish. To enhance the performance, a variety of plans may be implemented.

Data Augmentation techniques may assist in enhancing recognition in deteriorating classes. Decorative feature design by cleaning MFCC, chroma, or spectral features could further enhance the distinction between akin emotions. Further, converting the model structure with deeper layers or incorporating a consideration mechanism may enhance the feature extraction and emotion distinction. Finally, incorporating class weighting could solve possible class imbalances, affirming increased model creation.

Generally, while the model presents well in certain emotions like neutral, Further tuning can improve its ability to differentiate between similar emotions like angry, fear, and happy.

5.2 Classification Report

	precision	recall	f1-score	support
angry	0.94	0.84	0.89	139
disgust	0.97	0.87	0.92	162
fear	0.92	0.89	0.90	138
happy	0.83	0.87	0.85	157
neutral	0.89	0.95	0.92	246
sad	0.88	0.91	0.89	163
surprise	0.93	0.94	0.93	147
accuracy			0.90	1152
macro avg	0.91	0.90	0.90	1152
weighted avg	0.90	0.90	0.90	1152

Figure 4: Classification Report

The classification report makes available a wide-ranging valuation of the model's performance in guessing emotions created on precision, recall, and F1-score metrics. The model has accomplished an overall accuracy of 90%, indicative of strong performance. Among the specific emotions, the model accomplishes extremely well in recognizing emotions like disgust, neutral, and surprise. The disgust session shows the uppermost precision of 0.97, indicating that as soon as the model predicts 'disgust', it is just about always correct. In the same way, the surprise class has durable scores with a 0.93 precision and 0.94 recall, resulting in a high F1 score of 0.93. The neutral class stands out with a 0.95 recall, suggesting that most definite 'neutral' trials were successfully recognized. (Figure 4)

There are some locations, however, where the model's accuracy is slightly lower. For example, the angry class is recalled with a value of 0.84, so some 'angry' samples were misclassified. Although its precision remains high at 0.94, so accurate predictions when the model predicts 'angry', the lower recall means there are some false negatives. The happy class, on the other hand, has the lowest precision (0.83) and F1 score (0.85) of any of the emotions, so the model performs worst at distinguishing 'happy' samples from the rest.

Briefly, the model correctly classifies most of the emotions with extremely high recall and accuracy, but the angry and happy classes need to be improved. The misclassifications can be minimized and the performance of the model can be improved with data augmentation, class weighting, and enhancing feature extraction with deeper networks or attention.

5.3 ROC Curve

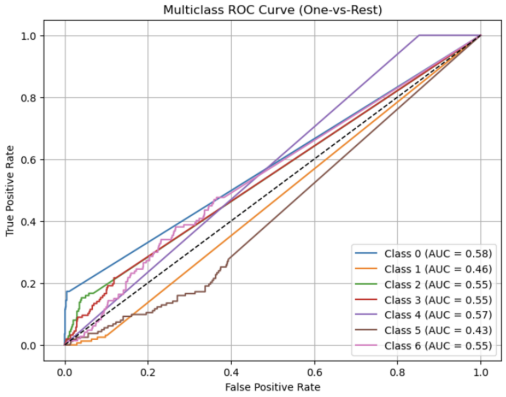

Figure 5: ROC Curve

The provided ROC curve shows the performance of the model for more than one class with the One-vs-Rest approach. There is a curve for every class, showing the relationship between the True Positive Rate (TPR) and False Positive Rate (FPR) at varying thresholds. The Area Under the Curve (AUC) measurements give an overview of the performance of each class, and higher AUC measurements indicate greater discriminative power.

Based on the plot, the model performs reasonably poorly in most classes. Class 0 and Class 4 perform best with AUC of 0.58 and 0.57, respectively, showing somewhat better-than-chance performance. Class 1 and Class 5 perform the worst with the worst AUC of 0.46 and 0.43, reflecting poorer-than-random-guess performance. (Fig. 5) Most other classes have an AUC of around 0.55, reflecting poor discriminability. The appearance of several curves close to plotting along the diagonal dash line (random guessing) suggests that the model is not doing well at distinguishing between classes.

To enhance the performance, methods like feature engineering, class balancing, and model optimization can be utilized. Augmentation of feature extraction techniques like MFCC, chroma, and Mel spectrogram can enable the model to learn more significant patterns from the data. Mitigation of possible class imbalances using oversampling or under-sampling techniques can enable the model to get proper training for all classes. Moreover, the optimization of the model architecture, hyperparameter tuning, or the addition of attention mechanisms could enhance its capacity to differentiate difficult emotions. Overall, although the existing model demonstrates average performance, some enhancements can greatly enhance its prediction capability.

5.4 Loss and Accuracy

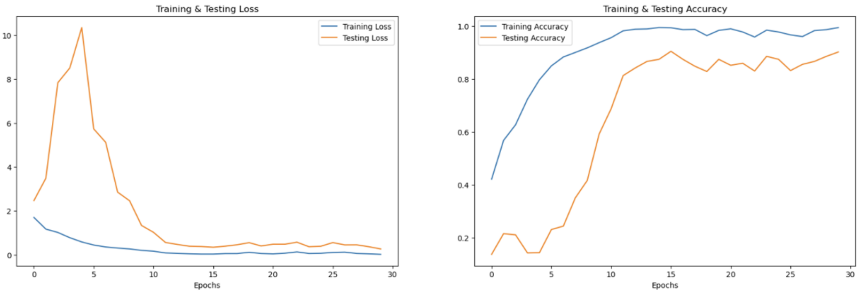

Figure 6: Loss 7 Accuracy

The plots provided show the training and testing performance of the model across 30 epochs. In the Training and Testing Loss plot, the training loss decreases consistently across the epochs, indicating that the model is learning effectively from the training set. The testing loss, however, sharply increases during the initial epochs, peaking around the 5th epoch, before gradually decreasing and stabilizing. This temporary spike shows that the model had a generalization problem in the initial epochs of training, potentially due to unstable weight updates. But with continued training, the testing loss converged to almost the training loss, which shows improved model performance. (Fig. 6)

In the training and Testing Accuracy plot, the training accuracy increases sharply and levels off a about 100% at epochs 15, whereas the testing accuracy increases slowly to about 85-90% by the training time. The large gap between the two lines represents overfitting when the model performs very well on the training set but is unable to reproduce that performance on new data.

To solve these issues, several techniques can be employes. Reducing the learning rate or applying learning rate decay can stabilize early training and prevent sudden increases in testing loss. Incorporating Dropout layers or L2 regularization can reduce overfitting by limiting the model's reliance on specific features. Moreover, applying early stopping can prevent over-training that leads to overfitting. Increasing the dataset with data augmentation can also improve generalization by exposing the model to more patterns of data. Applying these techniques, the overall model performance and stability can significantly be improved.

6 Conclusion

In sum, the Speech Emotion Recognition (SER) system shows a good pipeline that successfully processes audio data at stages ranging from data analysis to preprocessing, feature extraction, and model training. The system utilizes Mel-Spectrograms and fundamental features such as MFCCs and chroma to well recognize emotional states like happy, sad, angry, and neutral. The use of convolutional layers for feature learning, as well as dropout layers to avoid overfitting, was critical to improving the model's performance. In addition, the inclusion of data augmentation procedures like pitch shift, adding noise, and time-stretch also enhanced the capacity of the model to generalize from a wide variety of audio samples.

The present work adds value to affective computing and artificial intelligence through the provision of an organized and efficient framework for the recognition of emotion from audio. Through the integration of state-of-the-art feature extraction methods with deep learning models such as Conv1D, the system can accurately detect complex patterns in audio. Adding real-time recording support further increases the applicability of the system to real-world use cases like virtual assistants, health monitoring, and customer service platforms. The combination closes the gap between theoretical audio analysis and real-world emotion detection systems.

Future research can pursue several exciting avenues. The use of transformer-based models could enhance the system's capacity to process complex sequential data better. Furthermore, the exploration of more sophisticated data augmentation methods could improve the model's resilience in hostile environments. Adding attention mechanisms would sharpen the model's attention on important audio chunks, enhancing its precision in capturing faint emotional hints. Adding the dataset with various languages, cultural differences, and environmental conditions would also increase the adaptability of the model. Lastly, implementing the SER system in real-world applications like smart assistants, mental health care systems, and call center operations would ensure its practical utility and scalability.

References

[1] Kerkeni, L. (2022). Automatic speech emotion recognition using machine learning. Retrieved from https://hal.science/hal-02432557/

[2] Ye, Z. (2023). Emotion recognition based on convolutional gated recurrent units with attention. Speech Communication, 95, 17-25. https://doi.org/10.1080/09540091.2023.2289833

[3] Wang, H. (2024). Research on deep learning-based speech emotion recognition system. International Journal of Computer Science and Information Technology, 16(3). Retrieved from https://wepub.org/index.php/IJCSIT/article/view/2249/2467

[4] Wani, T. M. (2021). A comprehensive review of speech emotion recognition. IEEE Xplore. https://ieeexplore.ieee.org/abstract/document/9383000

[5] Jo, A. H. (2023). Speech emotion recognition based on two-stream deep learning model using Korean audio information. Electronics, 13(4), 2167. https://doi.org/10.3390/electronics13042167

[6] Al Dujaili, M. J. (2020). Speech emotion recognition based on SVM and KNN classifications fusion. ResearchGate.

[7] Sepeda, J. C. (2024). An enhancement of gated recurrent unit (GRU) for speech emotion recognition in the implementation of voice-based danger recognition system. ResearchGate.

[8] Sha, M. (2024). Speech emotion recognition using RA-GMLP model on time-frequency domain features extracted by TFCM. MDPI Electronics, 13(3), 588. https://doi.org/10.3390/electronics13030588

[9] Zhao, H. (2024). Knowledge enhancement for speech emotion recognition via multi-level acoustic feature. Speech Communication, 98, 32-40. https://doi.org/10.1080/09540091.2024.2312103

[10] Gupta, V. (2020). Emotion recognition of audio/speech data using deep learning approaches. Speech Communication, 119, 44-56. https://doi.org/10.1080/02522667.2020.1809089

[11] Raja, K. S. (2024). Speech emotion recognition using machine learning. ResearchGate. Retrieved from https://www.researchgate.net/publication/381165805_Speech_Emotion_Recognition_Using_Machine_Learning

[12] Singh, A. (2024). Analyzing the recent advancements for speech emotion recognition using machine learning techniques. ResearchGate.

[13] Liu, W. (2024). Development of a dataset for speech emotion recognition and analysis. e-speech: Development of a dataset for speech emotion recognition and analysis. Retrieved from [PDF link]

[14] Bou Nassif, A. (2019). Speech recognition using deep neural networks: A systematic review. IEEE Xplore. https://ieeexplore.ieee.org/document/8632885

[15] Durai Raj Vincent, P. M. (2025). Exploring deep learning methods for audio speech emotion detection: An ensemble of MFCCs, CNNs, and LSTM. Natural Sciences Publishing.

[16] Blal Er, M. (2020). A novel approach for classification of speech emotions based on deep and acoustic features. IEEE Xplore. https://ieeexplore.ieee.org/document/9285237

[17] Peng, Z. (2021). Efficient speech motion recognition using multi-scale CNN and attention. arXiv. https://export.arxiv.org/pdf/2106.04133v1.pdf

[18] Madanian, S. (2023). Speech emotion recognition using machine learning — A systematic review. ScienceDirect. https://www.sciencedirect.com/science/article/pii/S2667305323000911#se0280

[19] Aggarwal, A. (2022). Two-way feature extraction for speech emotion recognition using deep learning. Semantic Scholar. Retrieved from https://www.semanticscholar.org/reader/7b231d2ccaa48c00239047476144f555ca360fc4

[20] Liu, G. (2023). Speech emotion recognition based on emotion perception. ResearchGate. Retrieved from https://www.researchgate.net/publication/ 358442105_English_speech_emotion_recognition_method_based_on_speech_recognition

BIOGRAPHIES OF AUTHORS

Samyuktha S

Effectively completed BSc. at Vijaya College in Jayanagar, and now pursuing MSc in Computer Science, specialization in Data Science at Mount Carmel College, Samyuktha continues to showcase an insatiable eagerness for learning and a steadfast commitment to academic excellence. With her tireless dedication, boundless curiosity, and passion for making a difference, she stands ready to leave a lasting, positive impact on both the academic world and the broader society. She can be contacted at email: samyuktha6665@gmail.com

Sarwath Unnisa

Sarwath Unnisa, affiliated with Mount Carmel College, has contributed to various research areas, including cloud computing, AI-driven healthcare, IoT security, and geospatial data ethics. Her work spans multiple publications, covering topics such as AI applications in medical diagnostics, industry-integrated IoT, and deep learning for decision-making. Her research focuses on emerging technologies and explores the intersection of artificial intelligence, machine learning, and ethical considerations in modern computing environments. She can be contacted at email: sarwath@mccblr.edu.in