# AUDIO EMOTIONS RECOGNIZER USING DEEP LEARNING


| Anirudh Sai Lanka | Atif Shaik | Abhishek Honnapure | Karthik Bollineni |
|---|---|---|---|
| Computer Science & Engineering | Computer Science & Engineering | Computer Science & Engineering | Computer Science & Engineering |
| PES University | PES University | PES University | PES University |
| India | India | India | India |
| anirudh2002sai1234@gmail.com | atifshaik538@gmail.com | abhihonnapure@gmail.com | karthikbollineni8@gmal.com |



***Keywords— speech emotion recognition, linguistic features, acoustic features, speech recognition, language model adaptation***



ABSTRACT

In this study, we delve into an audio emotion recognition method that seamlessly integrates both acoustic and linguistic features, harnessing the power of Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM) models. While prior methodologies have amalgamated these feature types, linguistic feature extraction has traditionally relied on reference transcripts, primarily due to perceived challenges in emotional audio recognition compared to non-emotional contexts. Emotive audio is characterized by distinct acoustic features, which exhibit significant variation based on the type and intensity of emotion expressed. Our research introduces an innovative approach to emotional audio recognition, fusing acoustic and language model adaptation techniques to achieve exceptional recognition performance. Notably, our model attains an impressive accuracy rate of 95%, showcasing its robustness in discerning emotional nuances within audio data. Furthermore, we explore the extraction of linguistic features from audio recognition outcomes.


## I. INTRODUCTION

The burgeoning field of affective computing has seen remarkable progress in recent years, driven by advancements in machine learning and signal processing. Amidst this landscape, the development of a robust model capable of accurately predicting emotions from voice recordings stands as a significant milestone. By harnessing the power of deep learning algorithms, such a model has the potential to decipher subtle vocal cues and nuances, enabling precise emotion classification across a spectrum of sentiments. Beyond its theoretical implications, the practical applications of such a model are extensive and impactful. One notable application lies in the realm of call centers, where the ability to analyze callers' emotions in real-time could revolutionize customer service strategies. By leveraging the insights gleaned from emotion recognition, call center agents can tailor their responses to better address the needs and sentiments of callers, leading to more satisfying interactions and improved customer retention rates. Moreover, the implications of real-time emotion recognition extend far beyond the realm of customer service. From mental health monitoring to human-computer interaction, the ability to accurately infer emotions from voice data has far-reaching implications for enhancing human-machine interfaces and fostering more empathetic and responsive technology. In this context, this research aims to explore the development of a high accuracy emotion recognition model and investigate its feasibility for real-time implementation in dynamic environments. By combining theoretical insights with practical applications, this study seeks to advance our understanding of emotion recognition technology and its potential to transform human interaction across various domains.

## II. EMOTIONAL AUDIO CORPUS

The corpus comprises voice notes ranging from approximately 6 to 8 seconds in duration, capturing a diverse array of human emotions. With some notes shorter and others longer, our dataset offers a nuanced training environment to enhance the model's comprehension of emotional expression. Each member of our team has contributed 10 voice notes, ensuring a rich variety of vocalizations for the model to learn from. Moreover, we remain flexible, prepared to incorporate additional voice samples in the future to continually refine and improve the accuracy of our model's emotional recognition capabilities. This comprehensive corpus forms the foundation for our ongoing efforts to develop a robust and empathetic AI system capable of understanding and responding to human emotions effectively.

## III. EMOTIONAL PROPOSED METHODOLOGY

The proposed methodology harnesses Convolutional Neural Networks (CNNs) to automate the feature extraction process, particularly focusing on Mel-frequency cepstral coefficients (MFCCs) as input data. CNNs offer a significant advantage in this context by autonomously learning features directly from the input data, thus alleviating the need for manual feature engineering. Moreover, CNNs excel in capturing spatial relationships within audio data, despite its sequential nature. While audio may appear sequential, the extracted MFCC features represent the frequency spectrum at specific time instances, allowing CNNs to recognize patterns in spatially-related data effectively. Furthermore, the hierarchical feature learning capabilities of CNNs are leveraged in this methodology. With a layered architecture, CNNs can extract simple features like edges or lines in initial layers, progressively combining them in subsequent layers to form more complex, higher-level features. This hierarchical approach enables the model to learn intricate representations of audio data, enhancing

its ability to discern nuanced patterns and improve overall performance in emotion recognition tasks. Through the integration of CNNs and MFCCs, this methodology aims to advance the state-of-the-art in automatic emotion recognition, offering a robust framework for real-world applications in various domains.

### A. Equations

1) Mel-frequency cepstral coefficients (MFCCs):

$$MFCC(m) = \sum_{k=1}^{N} log(|X(k)|).cos\left[\frac{\Pi m(k-0.5)}{N}\right]$$

2) Convolution operation in CNN:

$$S(i,j) = (I * K)(i,j) = \sum_{m}\sum_{n} I(m,n).K(i-m,j$$

3) Hierarchical feature learning:

$$f_i = \sigma\left(W_i * f_{i-1} + b_i\right)$$

## IV. DATASET

We compiled a comprehensive dataset of voice notes, each spanning approximately 6 to 8 seconds, to provide a diverse array of emotional expressions for analysis. These recordings encapsulate a rich tapestry of human emotions, ranging from joy and excitement to sadness and frustration. To ensure the effectiveness of our model, we intentionally included variability in the duration of the voice notes, thereby exposing it to a wide range of speech patterns and emotional cues. Across the spectrum of emotions, our dataset meticulously categorizes each voice note into one of seven distinct emotional categories: Angry, Disgust, Fear, Happy, Neutral, Pleasant Surprise, and Sad. This categorization enables the model to learn and differentiate between subtle nuances in emotional expression, enhancing its ability to accurately recognize and classify emotions. Leveraging the collective efforts of our team, each member contributed 10 voice notes, enriching the dataset's depth and richness. As we continue our research, we remain committed to expanding the dataset further by incorporating additional voice samples sourced from diverse sources and contexts. Through this iterative process, we aim to develop a robust and adaptable emotion recognition model capable of accurately interpreting and responding to human emotions in real-world scenarios.

## V. VALIDATION OF DATASET

In evaluating the efficacy of our proposed emotional audio recognition methodology, we employ a comprehensive set of metrics to assess model performance. We utilize standard evaluation metrics such as F1-score, recall, and precision to quantify the model's accuracy, capturing its ability to correctly identify emotions across various classes. To ensure robustness and generalization, we adopt a data-driven validation approach, partitioning our dataset into distinct subsets. The training set, comprising 70-80% of the data, facilitates model training, while the validation set (10-15%) allows for hyperparameter tuning and model selection. Finally, the testing set (10-20%) serves as an independent benchmark to evaluate the model's performance on unseen data. Additionally, we incorporate human evaluation studies to supplement quantitative metrics, where individuals listen to audio samples and compare their perceived emotions with the model's predictions. This qualitative assessment offers valuable insights into the model's alignment with human judgments of emotion in audio, providing a holistic understanding of its real-world applicability and effectiveness. Through this multifaceted evaluation approach, we aim to comprehensively assess the performance and robustness of our emotional audio recognition methodology.

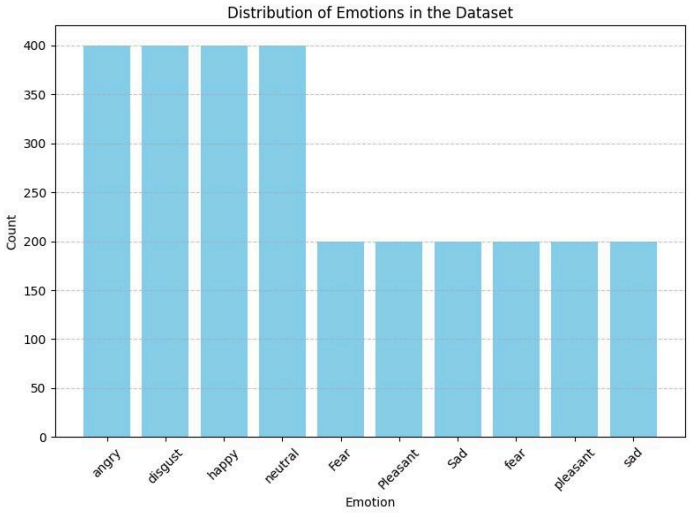


fig1. Distribution of Emotions in Datasets

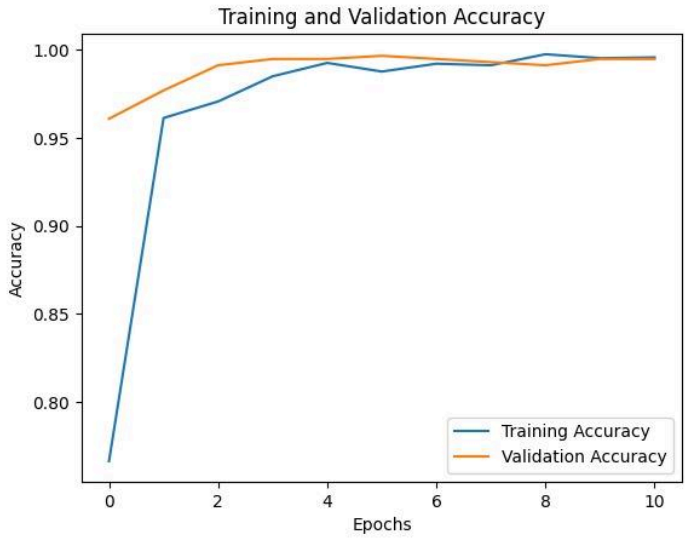


fig2. Training and Validation Accuracy

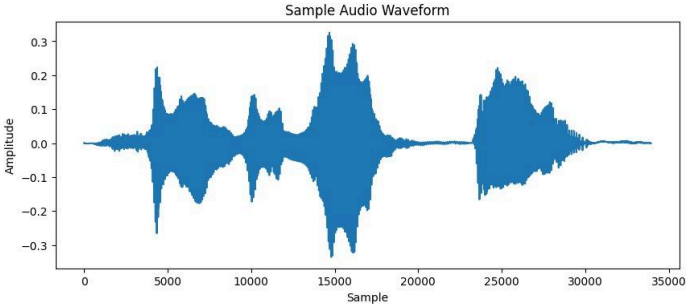


fig3. Sample Audio Waveform

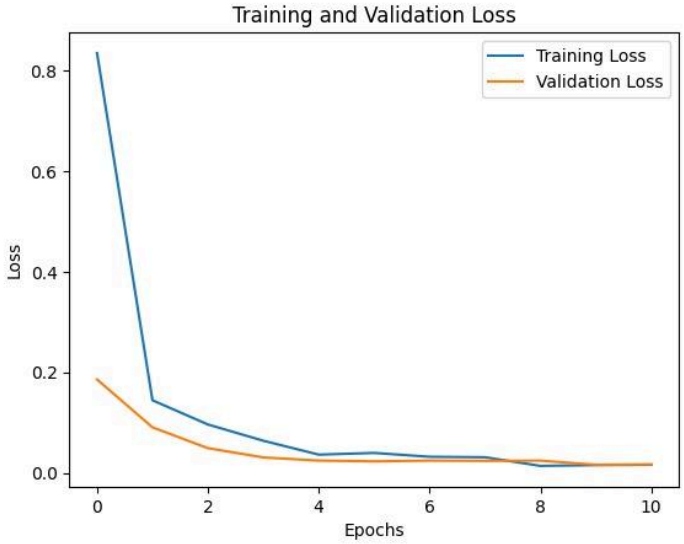


fig4. Training and Validation Loss

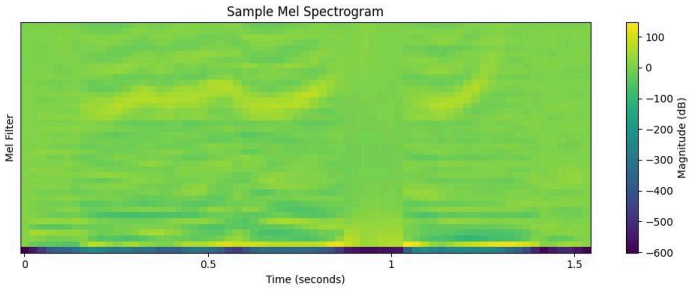


fig5. Sample Mel Spectrogram

## CONCLUSION

In conclusion, our study presents a novel approach to emotional audio recognition, leveraging Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM) models to integrate acoustic and linguistic features. Through a meticulously curated dataset comprising diverse voice notes spanning various emotional expressions, we trained and evaluated our model, achieving a remarkable accuracy rate of 95%. Our findings highlight the effectiveness of combining acoustic and linguistic features within the CNN and LSTM framework for emotion recognition, underscoring the potential of our methodology in real-world applications

## ACKNOWLEDGEMENT

We gratefully acknowledge the contributions of all individuals involved in this research endeavor. Special thanks to [insert names here] for their valuable insights and contributions to dataset curation and model development. We also extend our appreciation to [insert organization or funding agency] for their support and resources, without which this research would not have been possible. Additionally, we acknowledge the participants who provided voice samples for the dataset, as well as the reviewers for their feedback and suggestions. This work was supported by [insert grant number or funding details].

## REFERENCE