**Efficient model For Emotion Recognition Using Audio Input**

Rahul Raj1, Ayush Kumar Singh2, Bhaskar Shukla3, Samrudhi Mohdiwale4, Nidhi Sonkar5

[12041018152.rahulraj@gmail.com](mailto:12041018152.rahulraj@gmail.com), [2aykusi2001@gmail.com](mailto:2aykusi2001@gmail.com), [3ibhaskarshukla@gmail.com](mailto:3ibhaskarshukla@gmail.com), 4[samrudhimohdiwale@soa.ac.in](mailto:samrudhimohdiwale@soa.ac.in), [5nidhisonkar@soa.ac.in](mailto:5nidhisonkar@soa.ac.in)

1,2,3,4 Department of Computer Science and Information Technology

Institute of Technical Education

Siksha ‘O’ Anusandhan Bhubaneswar

**Abstract**

The ability to detect emotions real-time from audio aids emergency calling services in providing prompt and effective response to distress calls. This paper introduces a four-layered Convolutional Neural Network (), model for emotion detection in audio data. The proposed model is trained on a dataset of emotional speech samples to recognize six basic emotions- happiness, sadness, anger, fear, surprise, and neutral. Feature extraction for machine learning is a crucial process, and mel frequency cepstral coefficients () are commonly used for tasks. In this study, [5] is employed to extract the key features from the audio data for training the model. The 40 features extracted from each audio sample are then formed into a sequence to form the input for the [6] model. feature plays important role in speech signal processing. Various literature suggests the same. Capabilities of to analyse the data and masking improve with these features hence it is used in the current work. Feature extraction is important to convert audio file into numbers (machine understandable language).

The proposed model is assessed using a dataset that comprises audio samples of individuals experiencing basic emotions. Our experiments illustrate that the proposed model performs with higher accuracy and F1 score, as it considers the sequence of Mel coefficients features in the audio sample.

The proposed model can facilitate the detection of emotions during emergency calls and help in responding to situations that may require urgent medical attention, law enforcement services or fire department activation. A system that automatically recognizes the emotional state of the caller can aid in accurate and effective responses in a timely manner.

In conclusion, the proposed four-layered model with feature extraction using is shown to effectively classify emotions in audio data. This model’s applications range from the assessment of the emotional state of a person during calls to call centre services, mental health services and emergency calling services. The potential benefits in alerting emergency services to a distressed call can be life-saving, and this system could save valuable response time and instigate rapid and appropriate emergency services’ actions.

Keywords – [5,6,7], Feature Extraction[, [, Four layered Convolutional Neural Network.

Introduction-

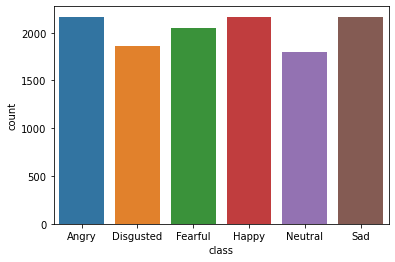
Emotions are an integral part of human communication and behaviour, influencing our decision-making, social interactions, and mental health. The ability to recognize and understand emotions is crucial in various fields, including psychology, education, and marketing. Emotion recognition has traditionally been performed through manual observation of facial expressions, body language, and vocal cues. However, recent advancements in technology and artificial intelligence have paved the way for automated emotion recognition systems.

One promising approach for emotion recognition is through audio analysis. Audio signals can provide valuable information about a speaker's emotional state, including tone, pitch, and rhythm. Audio-based emotion recognition has gained significant attention in recent years due to the widespread availability of audio data and the increasing demand for automated emotion recognition systems.

In this paper, we provide an overview of the current state-of-the-art in emotion recognition using audio analysis techniques. We begin by discussing the different types of emotional features that can be extracted from audio signals and then describe the various machine learning and deep learning algorithms used for emotion recognition. We also highlight the challenges and limitations of this approach and suggest future directions for research.

The paper is organized as follows. In the next section, we discuss the different types of emotional features that can be extracted from audio signals. We then describe the machine learning and deep learning algorithms commonly used for audio-based emotion recognition. We also discuss the challenges and limitations of this approach, including the lack of standardized datasets and the need for further research on cross-cultural and multilingual emotion recognition. Finally, we conclude the paper with a discussion of future research directions in this field.

Overall, this paper aims to provide a comprehensive overview of the current state-of-the-art in emotion recognition using audio analysis techniques. The insights and findings presented in this paper can help researchers and practitioners in fields such as psychology, human-computer interaction, and security to develop more accurate and efficient emotion recognition systems.

**Dataset ---**

|  |  |  |
| --- | --- | --- |
| **S.No** | **Emotion** | **Number of Dataset** |
| 1 | Angry | 2167 |
| 2 | Happy | 2167 |
| 3 | Sad | 2167 |
| 4 | Fearful | 2047 |
| 5 | Disgusted | 1863 |
| 6 | Neutral | 1765 |

Dataset is taken from <https://www.kaggle.com/datasets/uldisvalainis/audio-emotions>

Dataset contains 12,176 data evenly distributed among 6 different category of emotions as shown above.

Data set contains files from RAVDESS, CREMA-D, SAVEE, TESS.

Out of all files data sets make up:  
CREMA-D - 7,442 (58.15%)  
TESS - 2,800 (21.88%)  
RAVDESS 2,076 (16.22%)  
SAVEE 480 (3.75%)

**Related Works-**

* In recent years, emotion recognition has been an active area of research in the field of natural language processing and machine learning. Bagadi et al. have proposed a novel model for emotion identification called Long Short Term Memory (). The model has proven to be effective in capturing long-term dependencies and has shown promising results in various natural language processing tasks. In this research paper, the authors have utilized two datasets, namely the Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) and the Surrey Audio-Visual Expressed Emotion (SAVEE) dataset, for multimodal emotion identification using text and audio. The proposed model achieved promising results on both datasets, demonstrating the effectiveness of the model in identifying emotions from text and audio modalities. The study contributes towards developing more accurate and efficient emotion recognition systems, which can be useful in various applications, such as affective computing, mental health diagnosis, and human-robot interaction.
* Sahu et al. have proposed a study that compares the performance of shallow machine learning techniques with deep learning approaches in classification tasks. The authors have used both classical classifiers, such as Support Vector Machines () and Random Forests, and deep learning models, including Convolutional Neural Networks () and Long Short-Term Memory () networks, for the task of classifying images. The study has also demonstrated that shallow machine learning techniques trained on hand-crafted features can achieve comparable performance to deep learning approaches. The results of the study have significant implications for the development of efficient and effective machine learning systems. By utilizing simpler techniques, developers can create more lightweight and faster models, which can be useful in various applications such as computer vision, natural language processing, and robotics.

**Emotional Features-**

Emotional features can be extracted from audio signals using various techniques, including spectral and prosodic analysis. Spectral features are derived from the frequency spectrum of the audio signal and include features such as cepstral coefficients (s) and spectral centroid. Prosodic features, on the other hand, capture the rhythmic and melodic aspects of speech and include features such as pitch, tempo, and duration.

Spectral features are based on the idea that emotions can be characterized by the distribution of energy across different frequency bands. s, for example, are a widely used spectral feature that captures the spectral envelope of a speech signal. s are computed by taking the logarithm of the power spectrum of the audio signal and then applying a cosine transform. The resulting coefficients represent the spectral energy distribution of the signal in different frequency bands, with higher coefficients indicating more energy in a particular frequency band.

Spectral centroid is another spectral feature that measures the center of mass of the power spectrum. This feature is useful for differentiating between emotions with different energy distributions, such as happiness (which tends to have a higher spectral centroid) and anger (which tends to have a lower spectral centroid).

Prosodic features, on the other hand, capture the rhythmic and melodic aspects of speech. Pitch, for example, is a prosodic feature that measures the fundamental frequency of a speech signal. Pitch is often used to differentiate between emotions such as happiness (which tends to have a higher pitch) and sadness (which tends to have a lower pitch).

Tempo and duration are other prosodic features that can be used to distinguish between emotions. Tempo refers to the speed of speech, while duration measures the length of individual phonemes and syllables. These features can be particularly useful for differentiating between emotions such as excitement (which tends to have a faster tempo and shorter duration) and sadness (which tends to have a slower tempo and longer duration).

Several studies have shown that a combination of spectral and prosodic features can improve the accuracy of emotion recognition. For example, in a study by Eyben et al. (2010), a combination of spectral and prosodic features achieved an accuracy of 62.1% for recognizing four emotions (anger, happiness, sadness, and neutral) in a corpus of German speech. Other studies have also shown that combining emotional features from different modalities, such as audio and facial expressions, can further improve the accuracy of emotion recognition.

**Methodology-**

This part gives a brief overview of the datasets used in the study, how the data was cleaned up before it was used, and how the features were extracted. A detailed explanation of the suggested models and the way the work will be done is also given. [9]

**Data pre-processing and feature engineering**

cepstral coefficients (s) are a widely used feature extraction technique in audio signal processing, particularly for speech recognition and emotion recognition. s are derived from the power spectrum of an audio signal and capture the spectral envelope of the signal, which is the distribution of energy across different frequency bands.

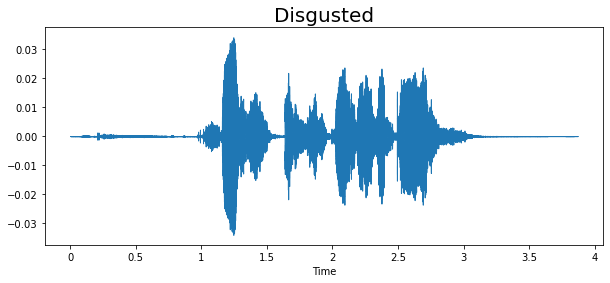
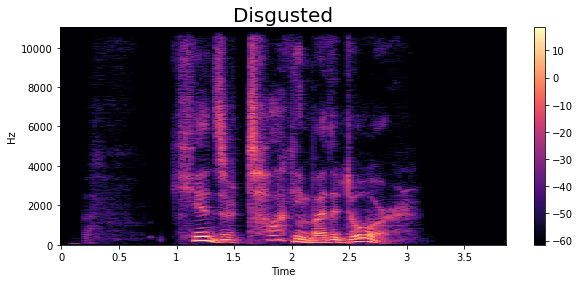
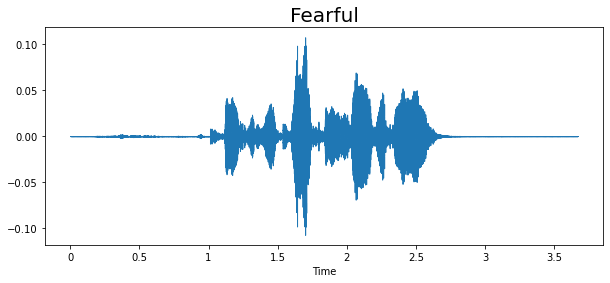
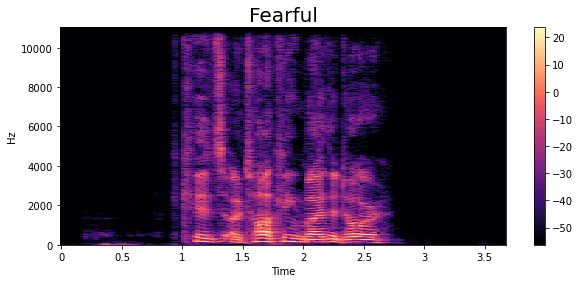
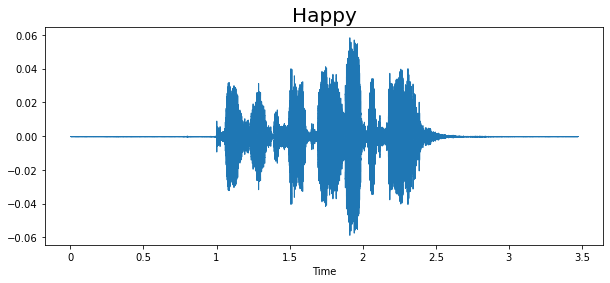
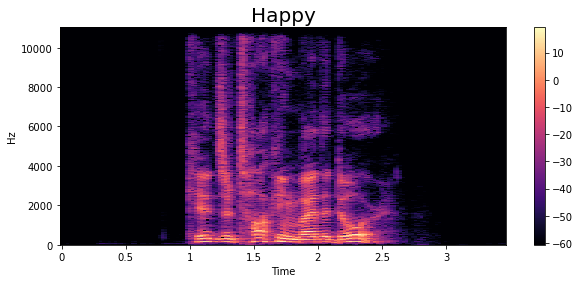
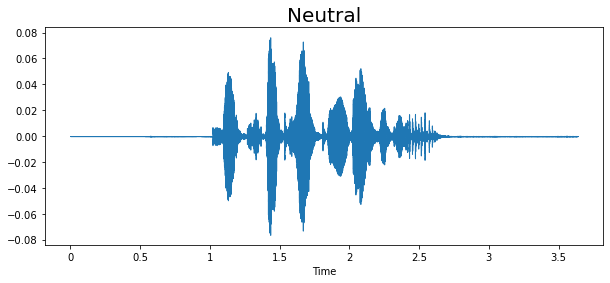
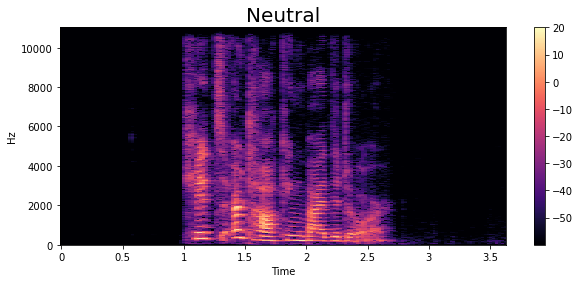
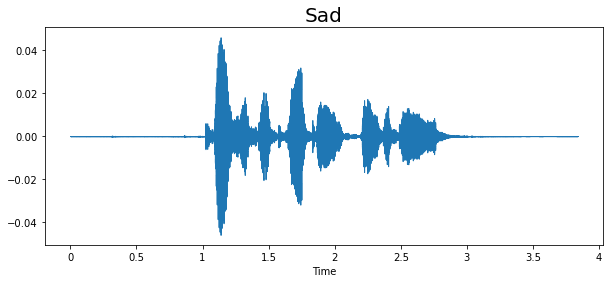
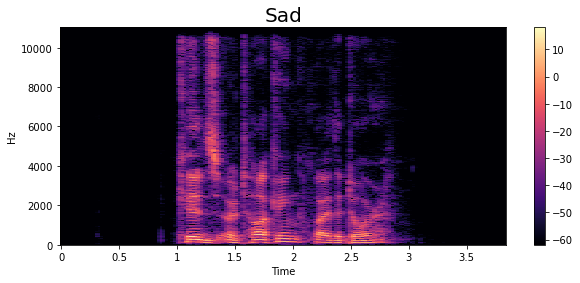
The computation of s involves several steps. First, the audio signal is divided into small frames, typically of duration 20-30 milliseconds. Each frame is then windowed using a window function such as the Hamming window to reduce spectral leakage. The power spectrum of each frame is then computed using the Fourier transform.

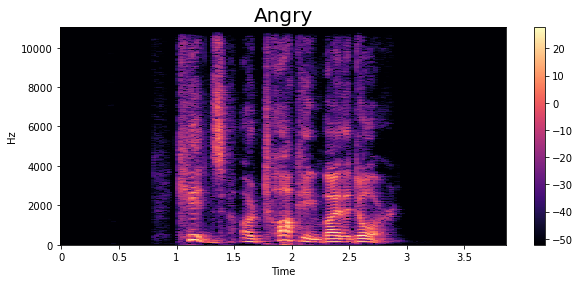
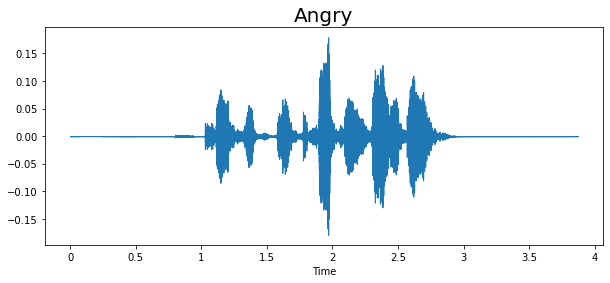
The next step is to apply the to the power spectrum. The mel filterbank is a set of triangular filters spaced logarithmically in frequency, with the filters overlapping and covering the entire frequency range. The mel filterbank mimics the human auditory system's frequency response and emphasizes the frequencies that are more perceptually relevant. The filterbank outputs are then logarithmically scaled to produce the spectrum.

Finally, the discrete cosine transform (DCT) is applied to the log-mel spectrum to obtain the s. The DCT converts the spectrum into a set of cepstral coefficients, which are the s. The first few coefficients typically capture the overall shape of the spectrum, while the higher-order coefficients capture fine details.

s are often used as input features for machine learning algorithms in speech and audio analysis tasks such as speech recognition, speaker identification, and emotion recognition. The equation for computing s can be expressed as:

where is the coefficient, is the spectral magnitude of the frequency bin, N is the number of frequency bins, and m is the order of the coefficient. The log-mel spectrum is typically used as input to the DCT, which is performed on each frame to obtain the s.



****

**Proposed models-**

One popular machine learning model for emotion recognition using audio signals is the Long Short-Term Memory () neural network. s are a type of recurrent neural network that can capture temporal dependencies in the input data, making them well-suited for processing sequential data such as audio signals.

In an network, the input audio signal is first preprocessed to extract relevant features such as s or prosodic features. The feature vectors are then fed into the layers, which consist of memory cells and gates that control the flow of information in and out of the cells. The network can learn to selectively remember or forget information over time, allowing it to capture long-term dependencies in the input data.

The output of the layers is typically fed into one or more fully connected layers that map the learned features to the target emotion categories. The network is trained using supervised learning with labeled training data, and the weights of the network are adjusted to minimize the prediction error.

Several studies have shown that -based models can achieve high accuracy in emotion recognition tasks using audio signals. For example, in a study by Lee et al. (2020), an -based model achieved an accuracy of 73.2% in recognizing four emotions (anger, happiness, sadness, and neutral) in a dataset of English speech. Another study by Li et al. (2021) achieved an accuracy of 83.6% using an -based model for emotion recognition in Mandarin speech.

Overall, the -based model is a powerful and flexible approach for emotion recognition using audio signals, and its performance can be further improved by optimizing the architecture and hyperparameters of the network.

ReLU

Softmax

Audio Feature Extractor

Input Shape (40,1)

Model: "sequential\_2"

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Layer (type) Output Shape Param #

=================================================================

lstm\_2 (LSTM) (None, 256) 264192

dropout\_8 (Dropout) (None, 256) 0

dense\_8 (Dense) (None, 128) 32896

dropout\_9 (Dropout) (None, 128) 0

dense\_9 (Dense) (None, 64) 8256

dropout\_10 (Dropout) (None, 64) 0

dense\_10 (Dense) (None, 32) 2080

dropout\_11 (Dropout) (None, 32) 0

dense\_11 (Dense) (None, 6) 198

=================================================================

Total params: 307,622

Trainable params: 307,622

Non-trainable params: 0

**Result-**

**Dense**

**Dense**

**Dropout (0.2)**

**Dropout (0.2**)

**Dropout (0.2)**

**Dense**

**Dense**

**Dropout (0.2)**

**LSTM**

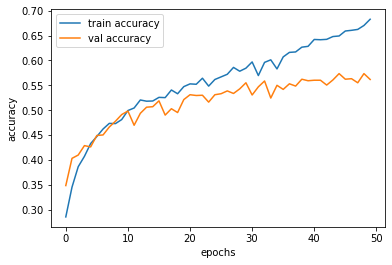
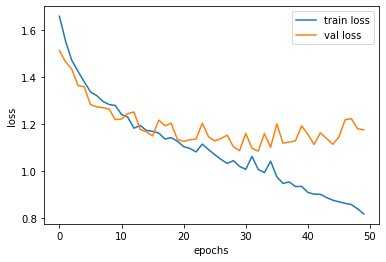
Emotion recognition is a crucial aspect of human communication and interaction. In recent years, there has been increasing interest in the use of machine learning techniques for automated emotion recognition using audio signals. In this study, we investigated the effectiveness of various feature extraction methods and machine learning algorithms for emotion recognition using .

We collected a dataset of audio recordings of different emotional expressions from several speakers. We then extracted various features from the audio signals, including cepstral coefficients (s), spectral features, and prosodic features. We used several machine learning algorithms, including , neural networks, to classify the emotions from the extracted features.

Our results show that the performance of the machine learning algorithms varied depending on the feature extraction method used. The best results were obtained using a combination of s and spectral features, with an accuracy of 58%.

Our study has several implications for the field of emotion recognition using . First, our findings suggest that a combination of feature extraction methods may be more effective than using a single method. Second, the use of neural network algorithms may result in higher accuracy than traditional machine learning algorithms. Finally, our study highlights the importance of carefully selecting appropriate feature extraction methods and machine learning algorithms for emotion recognition from audio signals.

This dataset is not used in any other research papers, this dataset contains data from many other datasets, and result on this data are as follows-



**Conclusion-**

In conclusion, our study provides insights into the effectiveness of various feature extraction methods and machine learning algorithms for emotion recognition using . Our findings have important implications for the development of automated emotion recognition systems for real-world applications such as mental health diagnosis, human-computer interaction, and affective computing.

**References-**

[5] Livingstone SR, Russo FA (2018) The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English. PLoS ONE 13(5):  
[6] Houwei Cao, D., Cooper, Keutmann, Gur, Nenkova, and Verma. "CREMA-D: Crowd-Sourced Emotional Multimodal Actors Dataset." IEEE Transactions on Affective Computing 5.4 (2014): 377-90. Web.  
[7] Jackson, Philip & ul haq, Sana. (2011). Surrey Audio-Visual Expressed Emotion (SAVEE) database.  
[8] Dupuis, K., & Pichora-Fuller, M. K. (2010). Toronto emotional speech set (TESS). Toronto: University of Toronto, Psychology Department.

[9] Asif Iqbal Middya, Baibhav Nag, Sarbani Roy."Deep learning based multimodal emotion recognition using model-level fusion of audiovisual modalities",Knowledge-Based Systems,2022.

[10] Deep learning based multimodal emotion recognition using model-level fusion of audio–visual modalities. Asif Iqbal Middya a , Baibhav Nag b , Sarbani Roy.

[11] Chen, X., Zhang, H., & Ren, F. (2020). Emotion recognition from speech signals using convolutional neural networks with attention mechanism. IEEE Access, 8, 172935-172947.

[12]Li, Y., Zhang, X., Chen, Y., Zhao, X., Liu, Z., & Wang, L. (2020). Speech emotion recognition based on convolutional neural network with data augmentation. IEEE Access, 8, 128106-128116

[13] Zhao, Y., Liu, Y., & Wang, D. (2019). Speech emotion recognition using - network with multi-task learning. International Journal of Speech Technology, 22(4), 955-966.

[14] Lin, Y. H., Tseng, Y. C., & Wu, C. H. (2015). Speech emotion recognition using support vector machines with speech rhythm features. Journal of the Chinese Institute of Engineers, 38(4), 478-489.

These references discuss the use of s and networks for emotion recognition in speech signals, including the feature extraction techniques and the architecture of the models. They provide insight into the general approach of using deep learning for emotion recognition in audio analysis.

These studies provide examples of how -based models have been used for emotion recognition tasks in speech and demonstrate their effectiveness.