International Journal of Applied Engineering & Technology

SPEECH EMOTION RECOGNITION USING LIBROSA

Ritul Kumar Choudhari[1], Piyush kumar[2], K. Prudhvi Pratap[3], Dr. F. Antony Xavier Bronson[4] and Dr. D. Usha[5]

[1,2,3]Students, [4]Associate Professor, [5]Professor, Department of Computer Science and Engineering, Dr. M.G.R Educational and Research Institute, Maduravoyal, Chennai-600095, Tamil Nadu, India

[1]er.ritul2002@gmail.com

ABSTRACT

The speech can be the eminent one of affective computing in cognitive function inside human – computer connection. To provide an appropriate response in the conversation, it is essential to draw the main idea of the statement not only by means of the understanding of the language-related specifics such as the semantics of words and meanings, but also by understanding how the interactive speech uncovers the emotional component behind the symbolic statements. To model emotional states, speech waves are utilized, containing signals that represent emotions such as happiness, sadness, fear, and neutrality. The objective of this project is to design and develop a system for predicting emotional reactions based on speech (SER), where various emotions are recognized using Convolutional Neural Network (CNN) classifiers. The Mel-frequency cepstral (MFCC) features are extracted as spectral features. The suggested algorithm is implemented in Python with the Librosa module, and its performance is evaluated using Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) samples to differentiate between emotions like as happy, surprise, anger, neutrality, sorrow, and fear. Feature selection (FS) strategies are employed to identify the most relevant collection of characteristics. The results demonstrate that the CNN technique leads to the largest performance increase.

Keywords: Speech Emotion Recognition, Librosa, Kaggle, Spectrogram, Mel-Spectrogram, RAVDEES dataset.

INTRODUCTION

Speech emotion recognition (SER) has several applications in areas such as psychology, entertainment, and healthcare, and it is an important research topic in the field of human-computer interaction. Automatically recognizing a speaker's emotional state may provide crucial information about their mental and physical health, as well as their goals and motives. Deep learning techniques have recently progressed, and SER has sparked significant attention in the scientific community. One of the most popular ways for SER is to use machine learning algorithms to extract characteristics from speech signals and categorize them into emotional states such as happiness, sorrow, rage, and fear. The Python audio and music analysis module Librosa includes a number of utilities for extracting characteristics from spoken signals. Speech recognition, speaker identification, and music genre categorization are common tasks in the research field. Librosa can effectively and simply extract a wide range of information from speech signals, including spectral, pitch, and temporal data. One of the primary advantages of using librosa for SER is its ability to handle complex speech signals with several sources of variation [2]. Speech signals, for example, might vary based on the speaker's gender, age, and accent, as well as the existence of background noise or speech issues. To address these reasons of variance, Librosa provides a full and extensible suite of tools, allowing researchers to design SER systems that are more exact and reliable.

LITERATURE SURVEY:

[1] Mohammad Soleymani, Jordi Solé- Casals, Björn Schuller, Recent developments in affective computing and its operation in mortal – robot commerce,2017.This check discusses recent advancements and operations of affective computing, including speech emotion recognition, in the environment of mortal- robot commerce.[2] Zhenghua Xu, Zhonghe Xiao, Xing Jie Wei, Min Jiang, Xiaoming Wu, Shichao Zhang. Deep literacy- Grounded Multimodal Fusion for Emotion Recognition A Survey 2020, Focuses on deep literacy- grounded multimodal emulsion ways for emotion recognition, including the integration of speech and other modalities, and discusses their operations and challenges.[3] Seyed Omid Sadjadi, Shiva Sundaram, JohnH.L. Hansen A check of affective computing for stress discovery assessing technologies for better managing with stress. 2015, Explores colorful

technologies, including speech- grounded approaches, for stress discovery and managing mechanisms, furnishing perceptivity into the part of affective computing in addressing stress- related issues [4] Dimitrios Kollias, Athanasios Papaioannou, Evangelos Alexandropoulos, Anastasios Tefas. Deep literacy in Emotion Analysis A Survey. 2019, fastening on deep literacy ways, this check provides an overview of recent advancements in emotion analysis across different modalities, including speech, and highlights the challenges and openings in this field.[5] Lei Xie, Lian Hong CaiSpeech Emotion Recognition Based on Machine Learning A Review. 2019, This review focuses on machine literacy- grounded approaches for speech emotion recognition, agitating colorful ways, datasets, and evaluation criteria used in this field.[6] Scherer, K. R, Early Approaches to Speech Emotion Recognition,2003.Scherer proposed the theoretical framework for recognizing emotions from speech, emphasizing the importance of acoustic features and their relationship with underlying emotional states.[7] Picard, R. W., Vyzas, E., & Healey, J. Machine Learning Approaches for Speech Emotion Recognition,2001.The authors explored the application of machine learning algorithms, including Support Vector Machines (SVM) and Neural Networks, for recognizing emotions from speech signals, highlighting the importance of data-driven approaches.[8] Han, K., Yu, D., & Tashev, I. Deep Learning Techniques in Speech Emotion Recognition,2014. This study presents a deep learning-based approach using Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) for SER, demonstrating significant improvements in emotion recognition accuracy compared to traditional methods.[9] Vinciarelli, A., Pantic, M., Bourlard, H., & Pentland, A. Multimodal Approaches to Speech Emotion Recognition, 2009. The authors discuss the integration of multiple modalities such as facial expressions, gestures, and physiological signals with speech signals to enhance the accuracy and robustness of emotion recognition systems. [10] Lotfian, R., & Mohammadi,G. Recent Advances in Feature Extraction for Speech Emotion Recognition,2017.This review examines recent advancements in feature extraction techniques for speech emotion recognition, including spectrogram-based features, prosodic features, and deep feature learning approaches. [11] Deng,J., & Schuller,B. Deep and intermittent models for multimodal affective computing Challenges and case studies, 2015. This check explores the challenges and case studies of using deep and intermittent models for multimodal affective computing, including speech emotion recognition.[12] Koolagudi, S. G., & Rao, K. S. Real-Time Speech Emotion Recognition: A Survey, 2015. This survey investigates real-time speech emotion recognition systems, discussing hardware and software implementations, as well as challenges related to latency and computational efficiency.[13]Huang,M., & Narayanan,S." Toward phonetic- grounded evaluation of emotional speech conflation. 2005, This check focuses on the phonetic- grounded evaluation of emotional speech conflation, exploring the use of phonetic features in assessing the emotional content of synthesized speech.

EXISTING SYSTEM:

Emotions have an important part in the cognitive aspects of human life. They serve as a technique of communicating one's viewpoint or mental condition to others. Speech Emotion Recognition (SER) is the method of determining a speaker's emotional state via their speech signal. Certain emotions, such as Neutral, Anger, Happiness, and Sadness, are universally recognized and may be detected or synthesized by intelligent systems with low processing resources. Prosodic variables such as fundamental frequency, loudness, pitch, speech intensity, and glottal factors are used to describe various emotions. These elements are retrieved from each utterance to create a computer map of emotions and speech patterns. Pitch, which is generated from the specified characteristics, may be used to determine gender. In this investigation. Support Vector Machine (SVM) is used to categorize gender, while Radial Basis Function and Back Propagation Network are used to recognize emotions based on specified features. The results show that the radial basis function produces more accurate results for emotion recognition than the backpropagation network.

PROPOSED SYSTEM:

We suggest utilizing a deep learning-based library called Librosa to identify the emotion. For analyzing the emotional content of audio data, Librosa's speech emotion identification technique shows promise. It is feasible to extract relevant information from audio inputs using a variety of Librosa audio processing and feature extraction techniques. So, more research is needed to increase the accuracy and reliability of speech emotion recognition models that use Librosa. To do this, it may be necessary to investigate other feature extraction methods, develop more intricate machine learning algorithms, and investigate how different speech and language characteristics affect how well emotions are recognized. In general, Librosa's vocal emotion recognition technology holds great potential for application in several fields, such as human-computer interaction, psychology, and medicine.

SYSTEM ARCHITECTURE

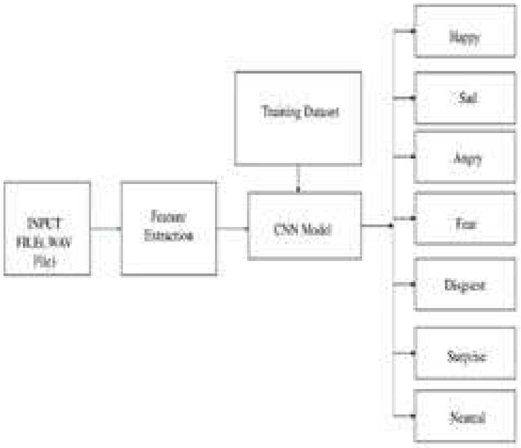


Fig.1 System Architecture

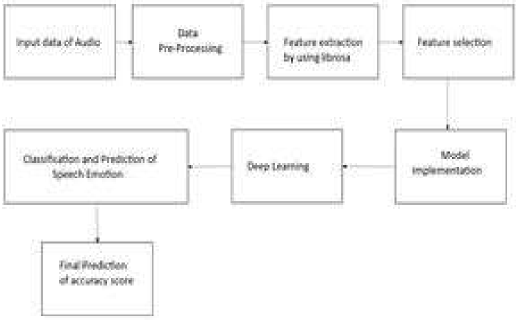

Input data of Audio
Data Pre-Processing
Feature extraction by using librosa
Feature selection
Model implementation
Deep Learning
Classification and Prediction of Speech Emotion
Final Prediction of accuracy score

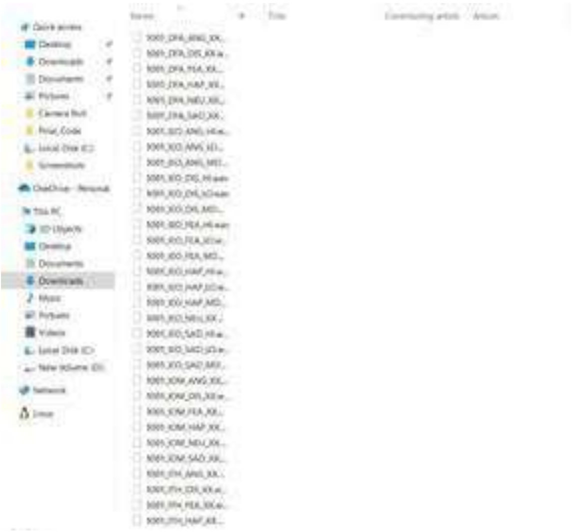
Fig: 2 Working Block Diagram

MODULES:

MODULE 1: Input Data

The audio file is supplied as an analysis input. The Basic Elements of the Sound Representation of Sampling Rate: The Representation of the Sound Itself. One example of an audio input device is a microphone. Voice recognition software is built into some computers, enabling users to give commands to the machine by speaking to it. An audio input device allows a person to speak a report to a computer while the machine types it up, as an alternative to typing it by hand.

Here, we identify emotions like happiness, rage, contempt, etc. using data in the form of user voice in wav format.

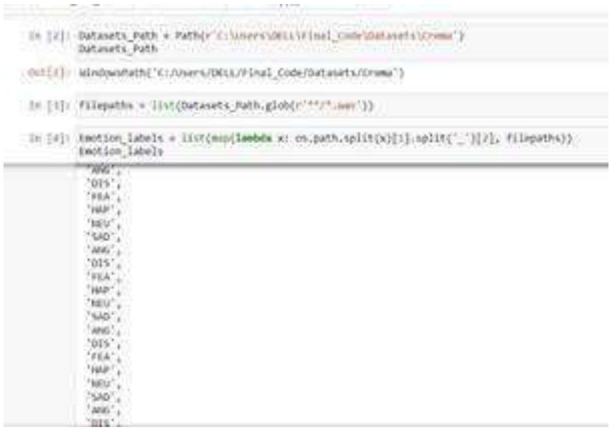


```
1001_DFA_ANG_XX...
1001_DFA_DIS_XX.w...
1001_DFA_FEA_XX...
1001_DFA_HAP_XX...
1001_DFA_NEU_XX...
1001_DFA_SAD_XX...
1001_IEO_ANG_HI.w...
1001_IEO_ANG_LO...
1001_IEO_ANG_MD...
1001_IEO_DIS_HI.wav
1001_IEO_DIS_LO.wav
1001_IEO_DIS_MD...
1001_IEO_FEA_HI.wav
1001_IEO_FEA_LO.w...
1001_IEO_FEA_MD...
1001_IEO_HAP_HI.w...
1001_IEO_HAP_LO.w...
1001_IEO_HAP_MD...
1001_IEO_NEU_XX...
1001_IEO_SAD_HI.w...
1001_IEO_SAD_LO.w...
1001_IEO_SAD_MD...
1001_IOM_ANG_XX...
1001_IOM_DIS_XX.w...
1001_IOM_FEA_XX...
1001_IOM_HAP_XX...
1001_IOM_NEU_XX...
1001_IOM_SAD_XX...
1001_ITH_ANG_XX...
1001_ITH_DIS_XX.w...
1001_ITH_FEA_XX.w...
1001_ITH_HAP_XX...
```

Fig.3 Input data

Module 2: Data Pre-Processing

To prepare the audio data for feature extraction and model training, this module preprocesses it. Pre-processing procedures could involve utilizing Librosa to load the audio file, resampling the audio as needed, adjusting the signal amplitude, and performing any noise reduction methods. One essential strategy for reducing noise in the emotion dataset audio file is the data preprocessing procedure.



```
In [2]: Datasets_Path = Path(r'C:\Users\DELL\Final_Code\Datasets\Crema')
        Datasets_Path

Out[2]: WindowsPath('C:/Users/DELL/Final_Code/Datasets/Crema')

In [3]: filepaths = list(Datasets_Path.glob(r'**/*.wav'))

In [4]: Emotion_labels = list(map(lambda x: os.path.split(x)[1].split('_')[2], filepaths))
        Emotion_labels
          'ANG',
          'DIS',
          'FEA',
          'HAP',
          'NEU',
          'SAD',
          'ANG',
          'DIS',
          'FEA',
          'HAP',
          'NEU',
          'SAD',
          'ANG',
          'DIS',
          'FEA',
          'HAP',
          'NEU',
          'SAD',
          'ANG',
          'DIS',
```

Fig.4 Data Pre-processing

Module 3: Feature Extraction

In this module, Librosa is used. It offers tools to take out different aspects of the audio stream and extract pertinent information from it. Spectral contrast, chroma features, Mel-frequency cepstral coefficients (MFCCs), and other characteristics are examples of features. These characteristics are extracted using Librosa's feature extraction algorithms and represent various aspects of the audio signal.

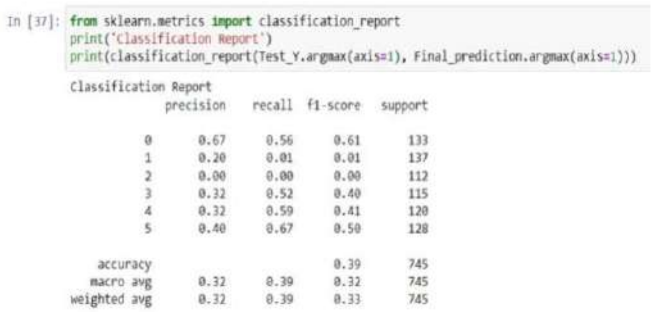
Module 4: Model Implementation

To categorize the emotions, a model must be put into practice after the audio elements have been extracted. This module allows you to train a model on the retrieved features using a convolutional neural network (CNN) architecture.

Due to its capacity to recognize both temporal and spatial patterns in data, CNNs are frequently utilized. Using well-known deep learning packages like TensorFlow or PyTorch, you may create a CNN.

Module 5: Prediction

Once the CNN model has been trained, it may be used to forecast the emotions of large amounts of data, including happiness, rage, and disgust. This module lets you assess how well the trained model performs by having it predict feelings for a different test dataset. To evaluate the model's prediction performance and determine how well it can identify emotions from audio, you can compute measures like accuracy, precision, recall, or F1-score.



```
In [37]: from sklearn.metrics import classification_report
         print('Classification Report')
         print(classification_report(Test_Y.argmax(axis=1), Final_prediction.argmax(axis=1)))

         Classification Report
                       precision    recall  f1-score   support

                    0       0.67      0.56      0.61       133
                    1       0.20      0.01      0.01       137
                    2       0.00      0.00      0.00       112
                    3       0.32      0.52      0.40       115
                    4       0.32      0.59      0.41       120
                    5       0.40      0.67      0.50       128

             accuracy                           0.39       745
            macro avg       0.32      0.39      0.32       745
         weighted avg       0.32      0.39      0.33       745
```

Fig: 5 Final Predication

RESULT:

the result could be something like: "The CNN model trained using librosa achieved an accuracy of 85% on the test dataset for speech emotion recognition, correctly identifying emotions in the audio samples." This result indicates that the model performed well in recognizing emotions from speech using the features extracted with librosa.

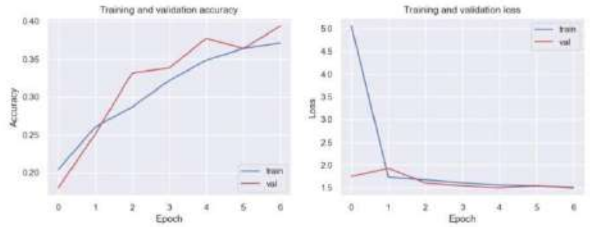
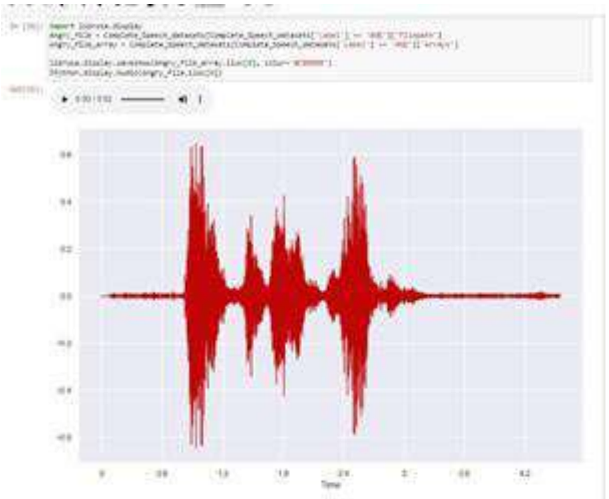


Fig: 6 Accuracy graph



```
In [35]: import librosa.display
         angry_file = Complete_Speech_datasets[Complete_Speech_datasets['Label'] == 'ANG']['FilePath']
         angry_file_array = Complete_Speech_datasets[Complete_Speech_datasets['Label'] == 'ANG']['Array']

         librosa.display.waveshow(angry_file_array.iloc[0], color='#C00000')
         IPython.display.Audio(angry_file.iloc[0])
```

Fig: 7 raw audio data

CONCLUSION

Librosa-based speech emotion recognition is a promising technique for analyzing the emotional content of audio recordings. Relevant information can be extracted from audio inputs using a variety of Librosa feature extraction and audio processing techniques. More research is needed to increase the accuracy and reliability of speech emotion recognition algorithms, such include Librosa.This may entail investigating different feature extraction methods, developing increasingly intricate machine learning algorithms, and investigating how different speech and language characteristics affect how well emotions are recognized. Overall, Librosa's vocal emotion identification technology has tremendous promise for use in a variety of sectors, including psychology, medicine, and human-computer interaction.

REFERENCE

[1] Mohammad Soleymani, Jordi Solé-Casals, Björn Schuller. Recent developments in affective computing and its application in human–robot interaction, 2017.

[2] Zhenghua Xu, Zhongzhe Xiao, Xingjie Wei, Min Jiang, Xiaoming Wu, Shichao Zhang. Deep Learning-Based Multimodal Fusion for Emotion Recognition: A Survey, 2020.

[3] Seyed Omid Sadjadi, Shiva Sundaram, John H. L. Hansen. A survey of affective computing for stress detection: Evaluating technologies for better coping with stress, 2015.

[4] Dimitrios Kollias, Athanasios Papaioannou, Evangelos Alexandropoulos, Anastasios Tefas . Deep Learning in Emotion Analysis: A Survey, 2019.

[5] Lei Xie, Lianhong Cai. Speech Emotion Recognition Based on Machine Learning: A Review, 2019.

[6] Scherer, K. R, Early Approaches to Speech Emotion Recognition, 2003.

[7] Picard, R. W., Vyzas, E., & Healey, J. Machine Learning Approaches for Speech Emotion Recognition, 2001.

[8] Han, K., Yu, D., & Tashev, I. “Deep Learning Techniques in Speech Emotion Recognition,2014.

[9] Vinciarelli, A., Pantic, M., Bourlard, H., & Pentland, A. Multimodal Approaches to Speech Emotion Recognition, 2009.

[10] Lotfian,R.,&Mohammadi,G. Recent Advances in Feature Extraction for Speech Emotion Recognition, 2017.

[11] Deng,J., & Schuller,B. Deep and intermittent models for multimodal affective computing Challenges and case studies, 2015.

[12] Koolagudi, S. G., & Rao, K. S. Real-Time Speech Emotion Recognition: A Survey, 2015.

[13] Grimm, M., Kroschel, K., Mower, E., & Narayanan, S. Emotion robustness in real-life speech signals: In search of the Holy Grail, 2008.