SPEECH EMOTION RECOGNIZATION USING AI

MINI PROJECT REPORT

SARANYA R 211720104129
SAMRAT K 211720104134
SIDDESH G K 21172010414

ABSTRACT

- As emotions play a vital role in communication, the detection and analysis of the same is of vital importance in today's digital world of remote communication. Emotion detection is a challenging task, because emotions are subjective. There is no common consensus on how to measure or categorise them. We define a SER system as a collection of methodologies that process and classifyspeech signals to detect emotions embedded in them. Such a system can find use in a wide variety of application areas like interactive voice based-assistant or caller-agent conversation analysis. In this study we attempt to detect underlying emotions in recorded speech by analysing the acoustic features of the audio data of recordings. Intelligent applications are interactive and require minimum user effort to function, and mostly function on voice-based input. This creates the necessity for these computer applications to completely comprehend human speech. A speech percept can reveal information about the speaker including gender, age, language, and emotion

- Several existing speech recognition systems used in IoT applications are integrated with an emotiondetection system in order to analyse the emotional state of the speaker. The performance of the emotion detection system can greatly influence the overall performance of the IoT application in many ways and can provide many advantages over the functionalities of these applications. This research presents a speech emotion detection system with improvements over an existing system in terms of data, feature selection, and methodology that aims at classifying speech percepts based on emotions, more accurately. This project is implemented in Kaggle platform using neural networks and deep learning techniques

1.INTRODUCTION

- Slide 1:

- Title: Speech Emotion Recognition

- 1. Importance:
- - Speech is the fastest and most natural form of human communication.
- - Researchers strive for effective computer-human interaction through speech recognition.

- 2. Emotional Speech Recognition:
- - Vital for natural human-computer interaction.
- - Applications: speech synthesis, customer service, education, forensics, and medical analysis.

- 3. Challenges:
- - Selecting effective features to differentiate emotions.
- - Language, accents, sentences, speaking style, and speaker characteristics pose difficulties.

- 4. Advancements:
- - Global hardware development for sound and speech recognition.
- - Significant progress in the '60s, supported by the US Department of Defense and DARPA.
- - Affordable and powerful voice recognition technology.

- 5. Voice as Dominant Interface:
- - AI advancements and abundant speech data suggest voice as the next dominant interface.

- 6. Methodology:
- - LSTM and RNN models employed.
- - Speech signals transformed into 2D representation using STFT.
- - CNNs and LSTM analyze the 2D representation.
- - Deep learning provides hierarchical representations.
- - Classifying results for each audio frame using probabilities.

2.PROBLEM STATEMENT

- 1. Project Aim:
- - Develop an AI system for accurately recognizing and classifying human emotions from speech signals.
- - Applications include human-computer interaction, virtual assistants, customer service, and mental health analysis.

- 2. Challenges to Address:
- - Accuracy: Building a highly accurate model to differentiate between emotional states.
- - Robustness: Ensuring the system performs well under various environmental conditions.
- - Real-time Performance: Designing an efficient algorithm for instant emotion recognition.
- - Generalization: Creating a model applicable across languages, cultures, and speech styles.
- - Data Availability: Collecting a comprehensive dataset of labeled speech samples.
- - Privacy and Ethics: Addressing privacy concerns and handling sensitive speech data responsibly.

3.OBJECTIVE

1.Objective:
•Develop an AI system for accurately detecting and classifying emotions in speech.
•Analyze audio recordings to identify the underlying emotional state conveyed by the speaker.
2.Applications:
•Human-Computer Interaction: Personalize AI responses based on user emotions.
•Customer Sentiment Analysis: Improve customer experiences by analyzing feedback.
•Mental Health Monitoring: Assess emotional well-being and detect potential disorders.
•Entertainment and Gaming: Enhance interactive experiences based on user emotions.
•Market Research: Understand consumer preferences and attitudes.
3.AI Model Development:
•Train machine learning algorithms on labeled emotional speech datasets.
•Extract relevant acoustic features and employ techniques like deep learning and pattern recognition.
•Enable real-time or near real-time emotion recognition.
.

4.DATASET SPECIFICATIONS

- CREMA-D Dataset:
- - Crowd Sourced Emotional Multimodal Actors Dataset
- - 7,442 original clips from 91 actors
- - 48 male and 43 female actors of diverse races and ethnicities
- - 12 sentences spoken with six emotions: Anger, Disgust, Fear, Happy, Neutral, Sad
- - Emotion levels: Low, Medium, High, Unspecified
- - Provides labeled data for studying multimodal expression and perception of basic acted emotions

- Purpose:
- - Generate standard emotional stimuli for neuroimaging studies
- - Varying degrees of length, intensity, and separation of visual and auditory presentation modalities
- - Group perceived emotion labels summarize multiple ratings for each clip

5)LITERATURE REVIEW

- - Title: Deep Learning Techniques for Speech Emotion Recognition, from Database to Models
- - Overview of commonly used databases for SER
- - Techniques to improve deep learning models for SER
- - Critical analysis of current research and future directions

- - Title: Clustering-Based Speech Emotion Recognition by Incorporating Learned Features and Deep BiLSTM
- - Clustering-based approach using learned features and deep BiLSTM
- - Feature learning and clustering-based classification
- - Promising method for improving SER accuracy

- - Title: A CNN-Assisted Enhanced Audio Signal Processing for Speech Emotion Recognition
- - Combination of audio signal processing and CNN for SER
- - Audio signal enhancement using spectral subtraction and mapping
- - CNN-based feature extraction and SVM classification

- - Title: Speech Emotion Recognition Based on LSTM Network with a Multi-Attention Mechanism
- - LSTM network with self-attention and guided attention mechanisms
- - New dataset for evaluation (KESD)
- - State-of-the-art performance on KESD dataset

- - Title: Probing Speech Emotion Recognition Transformers for Linguistic Knowledge
- - Investigating linguistic information exploitation in SER transformers
- - Probing processes based on publicly available tools
- - Valence predictions reactive to sentiment content, but not to intensifiers or reducers

5.2 Comparative study on various papers

- Methodology or Techniques used:
- - Deep learning architecture
- - Deep neural networks
- - Convolutional neural network (CNN)
- - Feedforward network
- - Long short-term memory (LSTM)

- Advantages:
- - Automatic detection of important features without human supervision

- .

- Issues:
- - Difficulty encoding the position and orientation of objects for CNNs
- - Challenges in encoding the position and orientation of pronunciations
- - Difficulty in classifying emotions with different positions

- Metrics used:
- - M1: Correctness of the input
- - M2: Layer segregation measures
- - M3: Accuracy of emotion classification

- .

6. PROPOSED MODEL

: Introduction

- - Speech Emotion Recognition and its potential applications
- - Proposed Model Architecture Overview

Model Architecture

- - Data Preprocessing: Audio Segmentation and Feature Extraction
- - Feature Representation: Temporal Context and Feature Normalization
- - Model Components: CNN, RNN, Attention Mechanism, Fusion Strategies

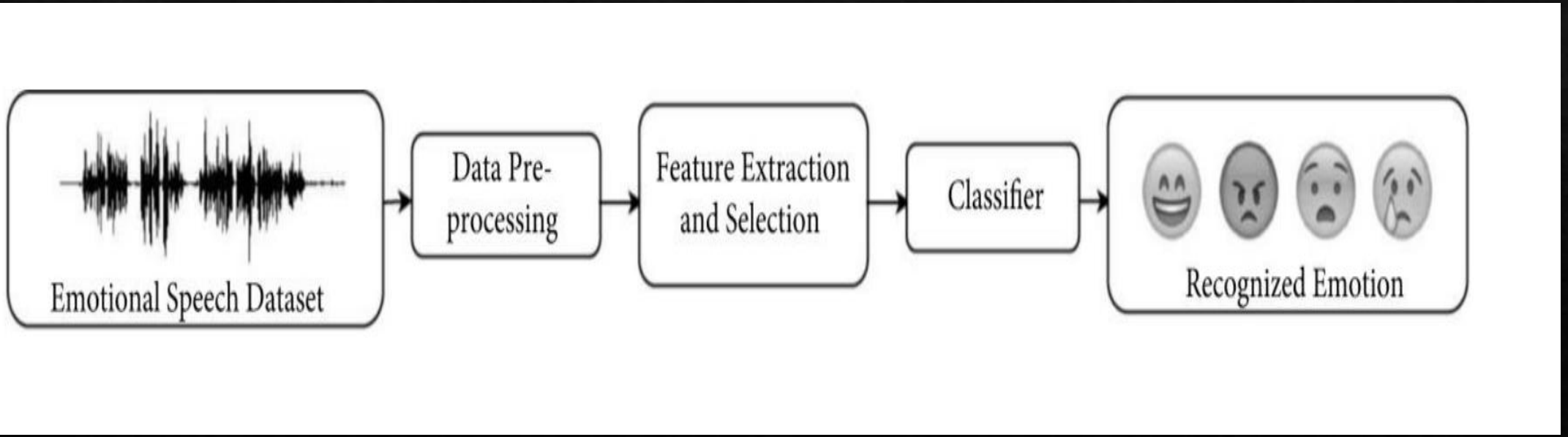
Training and Optimization

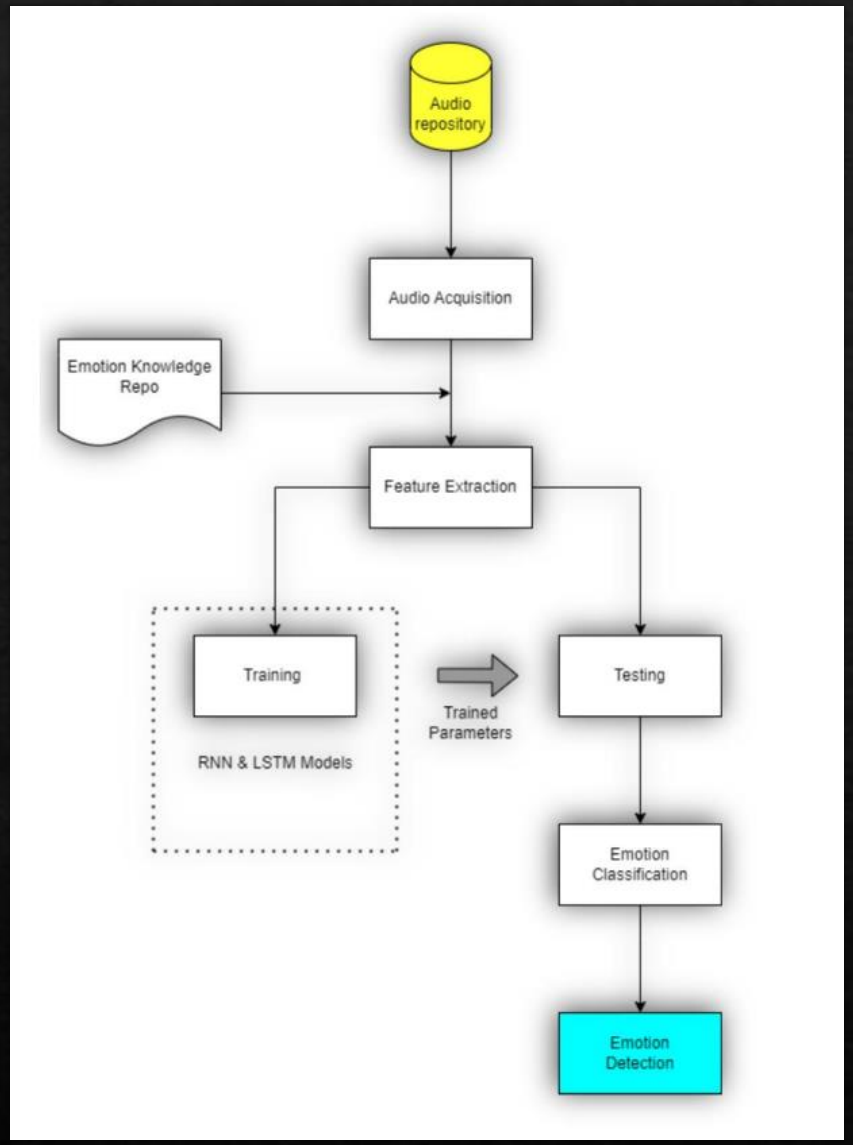
- - Dataset and Loss Function
- - Regularization Techniques and Optimization Algorithms
- - Hyperparameter Tuning

Evaluation and Deployment

- - Performance Metrics and Cross-Validation
- - Real-time Inference and Platform Considerations
- - Further Enhancements: Transfer Learning, Data Augmentation, Ensemble Methods

7.SYSTEM ARCHITECTURE




Audio repository
Audio Acquisition
Emotion Knowledge Repo
Feature Extraction
Training
RNN & LSTM Models
Trained Parameters
Testing
Emotion Classification
Emotion Detection

7.2Detailed Description of modules

- Introduction
- - Overview of the project and dataset (CREMA-D)
- - Importance of data preprocessing and visualization

- Data Preprocessing and Visualization
- - Importing necessary libraries (NumPy, Pandas, Seaborn, Matplotlib, sklearn, IPython, Tensorflow)
- - Loading the dataset and creating arrays for emotion paths and labels
- - Data visualization using Seaborn: Count plot and wave plot

Feature Extraction: MFCC

- - Creating a directory to map emotions with numerical values
- - Extracting MFCC features from audio files using a loop
- - Adding padding to extracted MFCC files using Keras

Data Splitting and Model Building

- - Splitting the data into training and testing datasets
- - Building a model with LSTM layers and a dense layer
- - Activation functions and optimizer selection

Model Training and Evaluation

- - Training the model using X_train and y_train with a batch size of 64
- - Tuning hyperparameters to improve model accuracy
- - Evaluation of the trained model

8.SOFTWARE REQUIREMENT SPECIFICATIONS

- 8.1Programming language: The language chosen for this project is Python 8.2Speech Processing Libraries: speech processing libraries such as librosa
- 8.3Machine Learning Frameworks: the speech emotions recognition needs a framework to recognise emotions from speech. The framework used here is Tensorflow
- 8.4Emotion Recognition Datasets: the dataset used here is Crema-D
- 8.5 Integrated Development Environment: Kaggle is an online IDE which has been used to code for the intended system

9 EXPERIMENTAL RESULTS AND DISCUSSION

- 9.1 SOURCE CODE:

- Introduction
- - Importance of importing libraries for data preprocessing and model building
- - Explanation of libraries used in the project

- Loading the Dataset
- - Importing necessary libraries (Pandas, NumPy, os, sys, librosa, sklearn, etc.)
- - Loading the CREMA-D dataset and creating arrays for emotion paths and labels
- - Creating data frames to store emotions and file paths

.

- MFCC Extraction
- - Mapping emotions to numerical values using a dictionary
- - Extracting MFCC features from audio files using librosa library
- - Padding the extracted MFCC features to ensure equal length

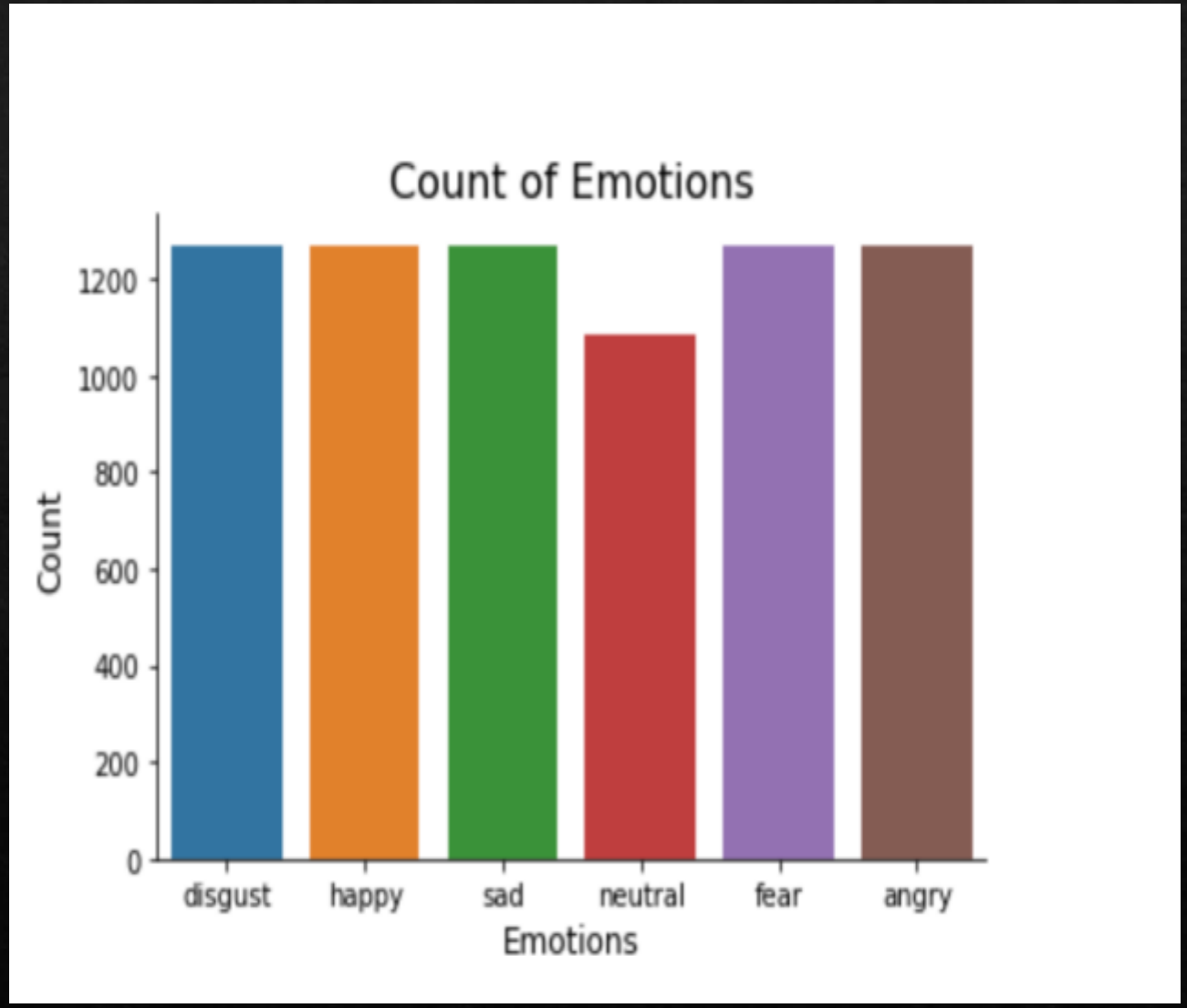
- Data Splitting and Model Building
- - Splitting the data into training, validation, and testing datasets
- - Defining the model architecture using LSTM and dense layers
- - Compiling the model with appropriate optimizer and loss function

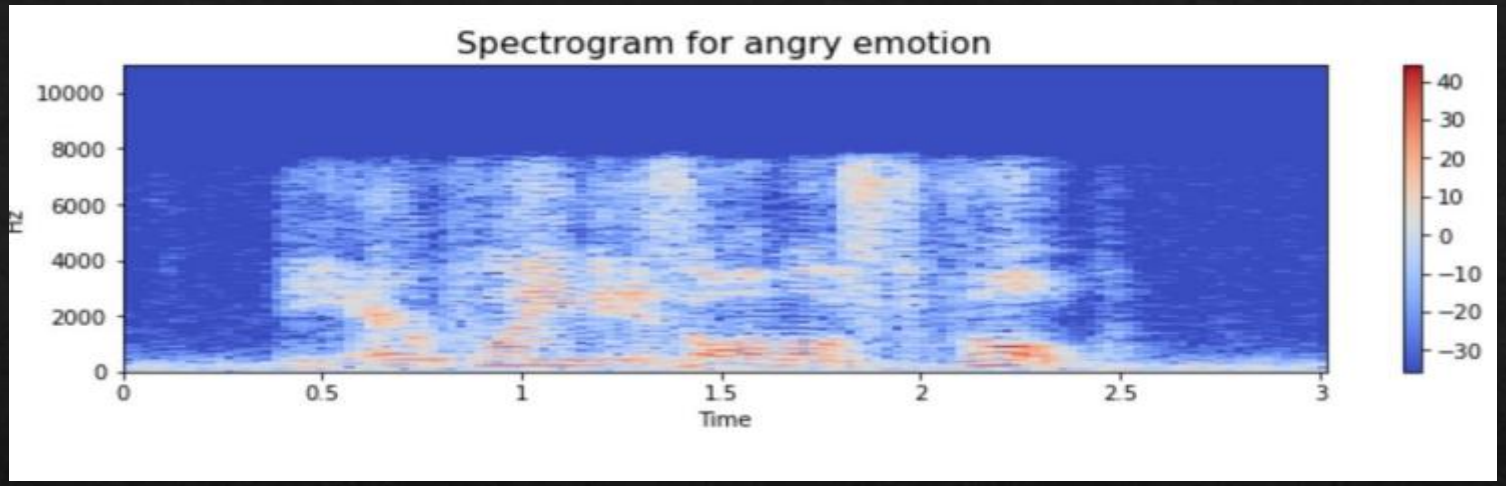
- Model Training
- - Training the model with the training data using the fit() function
- - Specifying batch size and number of epochs for training

- Model Evaluation
- - Evaluating the trained model on the testing dataset
- - Calculating the test accuracy of the model

9.2 SCREENSHOTS OF OUTPUTS:

	Emotions	Path
0	disgust	/kaggle/input/cremad/AudioWAV/1028_TSI_DIS_XX.wav
1	happy	/kaggle/input/cremad/AudioWAV/1075_IEO_HAP_LO.wav
2	happy	/kaggle/input/cremad/AudioWAV/1084_ITS_HAP_XX.wav
3	disgust	/kaggle/input/cremad/AudioWAV/1067_IWW_DIS_XX.wav
4	disgust	/kaggle/input/cremad/AudioWAV/1066_TIE_DIS_XX.wav


Count of Emotions
Count
1200
1000
800
600
400
200
0
disgust
happy
sad
neutral
fear
angry
Emotions




```
Model: "sequential"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
lstm (LSTM)                  (None, None, 124)         68448
_________________________________________________________________
lstm_1 (LSTM)                (None, 64)                48384
_________________________________________________________________
dense (Dense)                (None, 64)                4160
_________________________________________________________________
dropout (Dropout)            (None, 64)                0
_________________________________________________________________
dense_1 (Dense)              (None, 6)                 390
=================================================================
Total params: 121,382
Trainable params: 121,382
Non-trainable params: 0
```

10.REFERENCES

- 1) https://doi.org/10.3390/s21041249
- 2) https://doi.org/10.3390/s20010183
- 3) https://www.researchgate.net/publication / 340946514_Attention-LSTMAttention_Model_for_Speech_Emotion_Recognition_and_Analysis_of_IEMOCAP_Da tabase
- 4) https://www.ijraset.com/research-paper/speech-emotion-trecognition- systemusing-recurrent-neural-network
- 5) https://doi.org/10.3390/s22062378a
- 6)https://doi.org/10.48550/arXiv.1910.08874
- 7) https://arxiv.org/abs/2204.00400 8) https://ieeexplore.ieee.org/document/907878


THANK
YOU