SPEECH EMOTION RECOGNIZATION USING AI

MINI PROJECT REPORT

Submitted by

SARANYA R	**211720104129**
SAMRAT K	**211720104134**
SIDDESH G K	**211720104140**

in partial fulfillment for the award of the degree of

BACHELOR OF ENGINEERING

IN

COMPUTER SCIENCE AND ENGINEERING

RAJALAKSHMI INSTITUTE OF TECHNOLOGY

ANNA UNIVERSITY: CHENNAI 600 025

APRIL - MAY 2023

BONAFIDE CERTIFICATE

Certified that this project report **"Speech Emotion Recognization", is the bonafide work of Saranya R(211720104129), Sathya Samrat Ashok Chakravarthy(211720104134) and Siddesh G K(211720104140)** who carried out the mini project work under my supervision.

SIGNATURE:
Dr.R.SARAVANAN,M.E.,Ph.D.,
HEAD OF THE DEPARTMENT,
Professor,
Department of Computer Science and Engineering,
Rajalakshmi Institute of Technology,
Kuthambakkam Post,
Chennai-600124

SIGNATURE:
Dr.SRIDHAR,M.Tech.,Ph.D
SUPERVISOR,
Assistant Professor,
Department of Computer Science and Engineering,
Rajalakshmi Institute of Technology,
Kuthambakkam Post,
Chennai-600124

The report of this project work submitted by the above students in partial fulfillment for the award of Bachelor of Engineering degree in COMPUTER SCIENCE AND ENGINEERING under Anna University was evaluated and confirmed to be the report about the work done by the above students.

The University viva – voce is held on : _____

INTERNAL EXAMINER **EXTERNAL EXAMINER**

ACKNOWLEDGEMENT

We express our sincere gratitude to our honorable Chairperson **Dr.(Mrs.)THANGAM MEGANATHAN, M.A., M.Phil., Ph.D.,** and Chairman **Thiru .S.MEGANATHAN, B.E., F.I.E.,** for their constant encouragementto do this mini project and also during the entire course period.

We thank our Principal **Dr. P. K. NAGARAJAN** and our Head of the department **Dr. R. SARAVANAN M.E., Ph.D,** for their valuable suggestions and guidance for the development and completion of this mini project.

Words fail to express our gratitude to our Mini Project coordinator **R. ARUN KUMAR M.Tech.,(Ph.D)** Internal guide **Dr.SRIDHAR M.E.,Ph.D.,** who took special interest in our mini project and gave their consistent support and guidance during all stages of this mini project.

Above all we thank our parents and family members for their constant support and encouragement for completing this mini project.

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.

TABLE OF CONTENTS

CHAPTER NO:	TITLE	PAGE NO.
	ABSTRACT	
1	**Introduction**	5
2	**Problem statement**	6
3	**Objective**	7
4	**Dataset Specification**	9
5	**Literature Survey**	10
	5.1 Review on Various Schemes	10
	5.2 Comparative study on various subtitles	16
6	**Proposed Model**	24
7	**System Architecture**	25
	7.1 Architecture Diagram/Flow diagram/ flowchart	26
	7.2 Detailed Description of modules	27
8	**Software Requirement Specifications**	28
9	**Experimental Results and discussions**	29
	9.1 Source code	29
	9.2 Screenshots with Explanation	33
10	**List of References**	38

1.INTRODUCTION

Speech is the fast and best normal way of communicating amongst human. This reality motivatemany researchers to consider speech signal as a quick and effective process to interact between computer and human. It means the computer should have enough knowledge to identify human voice and speech. Although, there is a significant improvement in speech recognition but still researcher are away from natural interplay between computer and human, since computer is notcapable of understanding human emotional state. The recognition of emotional speech aims to recognise the emotional condition of individual utterer by applying his/her voice automatically Speech emotion recognition is mostly beneficial for applications, which need human-computer interaction such as speech synthesis, customer service, education, forensics and medical analysis . Recognising emotional conditions in speech signals are so challengeable area for several reason. First issue of all speech emotional methods is selecting the best features, which is powerful enoughto distinguish between different emotions. The presence of various language, accent, sentences, speaking style, speakers also add another difficulty because these characteristics directly change most of the extracted features include pitch, energy. Across the globe other nations developed hardware that could recognize sound and speech. And by the end of the '60s, the technology could support words with four vowels and nine consonants. Speech recognition made several meaningfuladvancements in this decade. This was mostly due to the US Department of Defense and DARPA. The technology to support voice applications is now both relatively inexpensive and powerful. With the advancements in artificial intelligence and the increasing amounts of speechdata that can be easily mined, it is very possible that voice becomes the next dominant interface. The method used to develop this project revolves around a model - LSTM and RNN. We transform the speech signal to 2D representation using Short Time Fourier Transform (STFT) after pre-processing. Then the 2D representation is analyzed through CNNs and Long ShortTerm Memory (LSTM) architectures. Deep learning involves hierarchical representations with increasing levels of abstraction. By

traversing sequentially constructed networks, the results corresponding to each selected audio frame are classified using a sum of probabilities.

2.PROBLEM STATEMENT

The aim of this project is to develop an Artificial Intelligence (AI) system capable of accurately recognizing and classifying human emotions from speech signals. Emotion recognition from speech plays a vital role in various applications such as human-computer interaction, virtual assistants, customer service, and mental health analysis. However, existing methods for speech emotion recognition often suffer from limitations in accuracy, robustness, and real-time performance.

The primary challenges to address include:

1. Accuracy: Developing a highly accurate emotion recognition model that can effectively differentiate between various emotional states, such as happiness, sadness, anger, fear, and neutrality, based on speech signals.
2. Robustness: Ensuring that the system remains robust and performs well under different environmental conditions, including background noise, variations in recording devices, and speaker characteristics.
3. Real-time Performance: Designing an efficient algorithm that can process speech signals in real-time, enabling applications that require immediate emotion recognition, such as live conversations and interactive systems.
4. Generalization: Creating a model that generalizes well across different languages, cultures, and speech styles, to ensure its applicability and effectiveness in diverse contexts.
5. Data Availability: Collecting and curating a comprehensive dataset of labeled speech samples covering a wide range of emotional states and diverse speaker demographics, which is crucial for training and evaluating the AI system.
6. Privacy and Ethics: Addressing privacy concerns and ensuring that the emotion recognition system respects user privacy by handling sensitive speech data responsibly

and securely.

By addressing these challenges, the proposed AI-based speech emotion recognition system will provide a valuable tool for improving human-computer interaction and enable various applications that rely on accurate and real-time understanding of human emotions from speech.

3.OBJECTIVE

The objective of speech emotion recognition using AI is to accurately detect and classify the emotional state or sentiment expressed in human speech. The goal is to develop an automated system that can analyze audio recordings of speech and identify the underlying emotions conveyed by the speaker.

Emotion recognition from speech can have various applications, including:

1. Human-Computer Interaction: By understanding the emotional state of users, AI systems can adapt their responses and interactions accordingly, providing more personalized and empathetic experiences.

2. Customer Sentiment Analysis: Emotion recognition can be used to analyze customer service calls, feedback, or social media posts to gauge customer satisfaction, identify potential issues, and improve the overall customer experience.

3. Mental Health Monitoring: Speech emotion recognition can aid in monitoring and assessing mental health conditions by analyzing speech patterns, providing insights into a person's emotional well-being and detecting potential signs of depression, anxiety, or other disorders.

4. Entertainment and Gaming: Emotion recognition can enhance interactive experiences in games, virtual reality, or augmented reality environments by adapting

the content or gameplay based on the user's emotional responses.

5. Market Research: Analyzing emotions expressed in focus groups, surveys, or interviews can help businesses gain a deeper understanding of consumer preferences, opinions, and attitudes towards products or services.

The primary objective is to develop AI models that can accurately and reliably classify emotional states such as happiness, sadness, anger, fear, surprise, or neutral sentiment from speech signals. This involves training machine learning algorithms on large datasets of labeled emotional speech recordings, extracting relevant acoustic features, and employing various techniques such as deep learning, pattern recognition, and signal processing to recognize and classify emotions in real-time or near real-time scenarios.

4.DATASET SPECIFICATIONS

The dataset used in this project is **CREMA-D - Crowd Sourced Emotional Multimodal Actors Dataset. CREMA-D** is a data set of 7,442 original clips from 91 actors. These clips were from 48 male and 43 female actors between the ages of 20 and 74 coming from a variety of races and ethnicities (African America, Asian, Caucasian, Hispanic, and Unspecified). Actors spoke from a selection of 12 sentences. The sentences were presented using one of six different emotions (Anger, Disgust, Fear, Happy, Neutral, and Sad) and four different emotion levels (Low, Medium, High, and Unspecified). a labeled data set for the study of multimodal expression and perception of basic acted emotions. The large set of expressions was collected as part of an effort to generate standard emotional stimuli for neuroimaging studies, and these require varying degrees of length and intensity and separation of visual and auditory presentation modalities. Multiple ratings for thesame clip are summarized in group perceived emotion labels.

5)LITERATURE REVIEW

5.1Review on Various Schemes

1. Deep Learning Techniques for Speech Emotion Recognition, from Database to Models

The review article examines the current state of research on speech emotion recognition (SER) using deep learning techniques. The article provides an overview of the most commonly used databases for SER, including the Berlin Emotional Speech Database (EMO-DB), the Interactive Emotional Dyadic Motion Capture (IEMOCAP) database, and the Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS), among others. The review also examines the various techniques used to improve the performance of deep learning models for SER, such as transfer learning, data augmentation, and feature selection. Finally, the article provides a critical analysis of the current state of research on deep learning techniques for SER and discusses the challenges and future directions for this field. Overall, the review provides a comprehensive overview of the current state of research on deep learning techniques for SER and is a valuable resource for researchers and practitioners working in this field.

2. Clustering-Based Speech Emotion Recognition by Incorporating Learned Features and Deep BiLSTM

In this paper, the authors propose a clustering-based approach to speech emotion recognition (SER) that incorporates learned features and a deep bidirectional long short- term memory (BiLSTM) network. The proposed approach consists of two main components: feature learning and clustering-based classification.

In the clustering-based classification phase, the authors use a clustering algorithm to

group similar speech segments together based on their learned feature representations. They then assign a label toeach cluster based on the majority emotion label of the speech segments in that cluster.

Overall, the proposed clustering-based approach to SER that incorporates learned features and a deep BiLSTM network is a promising method for improving the accuracy of SER systems. The approach is flexible, scalable, and has the potential to be applied to other domains beyond speechemotion recognition.

3.A CNN-Assisted Enhanced Audio Signal Processing for Speech Emotion Recognition

In this paper, the authors propose a method for speech emotion recognition (SER) that combines audio signal processing techniques with a convolutional neural network (CNN) to improve the accuracy of emotion classification. The proposed method consists of two main components:

A) audio signal enhancement

B) feature extraction using a CNN.

In the audio signal enhancement phase, the authors use a combination of spectral subtraction and spectral mapping techniques to reduce noise and improve the signal-to- noise ratio of speech signals. This enhances the quality of the input signals and helps to improve the accuracy of subsequent processing. In the feature extraction phase, the authors use a CNN to extract high-level features from the enhanced speech signals. The CNN is trained on a large-scale dataset to capture discriminative features that are important for emotion recognition. The extracted features are then used to classify the emotion of the speech signal using a support vector machine (SVM) classifier. Overall, the proposed method is a promising approach for improving the accuracy of SER systems by combining audio signal processing techniques with a CNN-based feature extraction approach. The method is flexible, scalable, and has the potential to be applied to other domains beyond speechemotion recognition

4. **Speech emotion recognition based on lstm network with a multi-attention mechanism**

Speech emotion recognition is a challenging task due to the complex nature of emotional expressions and the variability of speech signals. In recent years, deep learning techniques, especially recurrent neural networks (RNNs), have shown promising results in this field. In this paper,we propose a new approach for speech emotion recognition based on a long short-term memory (LSTM) network with a multi-attention mechanism. The proposed model utilizes both self- attention and guided attention mechanisms to selectively attend to relevant segments of the input sequence. The self-attention mechanism allows the model to learn the dependencies between different time steps in the input sequence, while the guided attention mechanism focuses on the

most informative frames for emotion classification. We also introduce a new dataset, called the Korean Emotional Speech Dataset (KESD), for evaluating the proposed method. Experimental results show that the proposed model achieves state-of-the-art performance on the KESD dataset and outperforms other existing methods on the benchmark datasets IEMOCAP and MSP- IMPROV. These results demonstrate the effectiveness of the proposed multi-attention mechanism for speech emotion recognition.

5.Probing speech emotion recognition transformers for linguistic knowledge

Large, pre-trained neural networks consisting of self-attention layers (transformers) have recently achieved state-of-the-art results on several speech emotion recognition (SER) datasets. These models are typically pre-trained in self-supervised manner with the goal to improve automatic speech recognition performance – and thus, to understand linguistic information. In this work, we investigate the extent in which this information is exploited during SER fine-tuning. The main contribution of this work relies on providing comprehensive, reproducible probing processes based on publicly available tools with an emphasis on the linguistic information learnt by SER models.

It is based on three probing methodologies like re-synthesising speech signals from automatic transcriptions using ESPnet Using a reproducible methodology based on open-source tools, we synthesise prosodically neutral speech utterances while varying the sentiment of the text. Valence predictions of the transformer model are very reactive to positive and negative sentiment content, aswell as negations, but not to intensifiers or reducers, while none of those linguistic features impact arousal or dominance. These findings show that transformers can successfully leverage linguistic information to improve their valence predictions, and that linguistic analysis should be included in their testing.

6.Speech Emotion Recognition System Using Recurrent Neural Network in Deep Learning

The speech signal is transformed into analogue and digital waveform which can be understood by the machine. Speech technologies are broadly used and seen to have unlimited uses. In many of the human-machine interface applications, emotion recognition from the speech signal is considered tobe the research topic for many years. For this purpose, for the identification of the emotions from the speech signal, many systems have been developed until now. In this paper, speech emotion recognition based on the previous technologies which use different models and methods for the emotion recognition is reviewed and a new approach is suggested.They are used to differentiate emotions such as anger, happiness, neutralstate, etc.

The intended system is going to be proposed such that it takes the input as speech both live and audio file and detects and recognizes the emotion behind that speech. After recognizing it, the output will be represented as the emotion in which the speech was spoken. There are various types of emotions included in this system such as happy, neutral, sad, etc. We have proposed to use the Recurrent Neural Network which is a part of Deep Learning Algorithms in order to increase our accuracy as compared to

others models and methods which are in existence. In RNN, one data pointor the current data depends upon the previous data point to perform an overall view. The model predicts the emotions based on the speech data provided during its execution. The conclusionis based on the applications which are fast emerging based usage of speech and vision systems.

Examples include effective evaluation, smart and quick system transportation and correct medicineprescriptions. This paper promise to deliver a comprehensive survey emphasising on the demands of speech and vision systems with the view of both hardware and software systems. The technologies which discussed in machine learning are fast gaining access and aim to revolutionise the areas of research and development in speech and vision systems.

7.Two-Way Feature Extraction for Speech Emotion Recognition Using Deep Learning

In this paper, they have explored two different methods of extracting features to address effective speech emotion recognition. Initially, two-way feature extraction is proposed by utilizing super convergence to extract two sets of potential features from the speech data. For the first set of features, principal component analysis (PCA) is applied to obtain the first feature set. Thereafter, a deep neural network (DNN) with dense and dropout layers is implemented. In the second approach, mel-spectrogram images are extracted from audio files, and the 2D images are given as input to the pre-trained VGG- 16 model. RAVDESS dataset provided significantly better accuracy than using numeric features on a DNN .Several feed forward activities and recurrent activities are conducted to get the desired output. The minimum and maximum frequency it attained are 40% and 82.35% respectively.

8. Robust Speech Emotion Recognition Using CNN+LSTM Based on Stochastic FractalSearch Optimization Algorithm

They proposed an optimized deep learning model in which the hyper parameters are optimized to find their best settings and thus achieve more recognition results. This deep learning model consistsof a convolutional neural network (CNN) composed of four local feature-learning blocks and a longshort-term memory (LSTM) layer for learning local and long-term correlations in the log Mel- spectrogram of the input speech samples. They used d stochastic fractal search (SFS)-guided whale optimization algorithm (WOA) to improve the performance of this deep network. Results show that the proposed system yields an average accuracy of 86% and the best accuracy of 99% in the four speech emotion datasets.

9. Speech emotion recognition with dual-sequence LSTM architecture

A new dual level model that predicts emotions based on both MFCC features and mel-spectrogramsproduced from raw audio signals is Proposed in this given paper. The dual-level architecture contains two separate models, MLSTM and MDS-LSTM, the first for the MFCC features and the second for the two melspectrograms. Each of these two models has a classification layer, the outputs of which are averaged to make the final prediction. The efficiency attained by the system proposed in this paper is on average, a weighted accuracy of 72.7% and an unweighted accuracy of 73.3%

10. Evaluating deep learning architectures for Speech Emotion Recognition - Haytham M. Fayek, , Margaret Lech , Lawrence Cavedon.

In this paper, A deep multi-layered neural network composed of several fully-connected, convolutional or recurrent layers ingests a target frame (solid line) concatenated with a number of context frames (dotted line) to predict the posterior class probabilities corresponding to the target frame. Several feed forward activities and recurrent activities areconducted to get the desired output. The minimum and maximum frequency it attained are 65% and 90% respectively.

11. Speech emotion recognition using hidden Markov models

It uses short time log frequency power coefficients (LFPC) to represent the speech signals and a discrete hidden Markov model (HMM) as the classifier. The emotions are classified into six categories. The category labels used are, the archetypal emotions of Anger, Disgust, Fear, Joy, Sadness and Surprise. Performance of the LFPC feature parameters is compared with that of the linear prediction Cepstral coefficients (LPCC) and mel-frequency Cepstral coefficients (MFCC) feature parameters commonly used in speech recognition systems. Results show that the proposed system yields an average accuracy of 78% and the best accuracy of 96% in the classification of six emotions.

12.Speech Emotion Recognition Using CNN - Zhengwei Huang , Ming Dong , QirongMao , Yongzhao Zhan.

Deep learning systems, such as Convolutional Neural Networks (CNNs), can infer a hierarchical representation of input data that facilitates categorization. Speech Emotion Recognition (SER) using semi-CNN is used in this given paper. The two training stages of semi-CNN is used in this paper.. In the first stage, unlabeled samples are used to learn candidate features by contractive convolutional neural network with reconstruction penalization. The candidate features, in the second step, are used as the input to semi- CNN to learn affect-salient, discriminative features using a novel objective function that encourages the feature saliency, orthogonality and discrimination. Pattern recognition methodology is used here in this paper. The minimum and maximum efficiency attained bythe system proposed in this paper is 71% and 92% respectively.

5.2 Comparative study on various papers

(Title, Year, Authors)	Methodology or Techniques used (Mention specific algorithms or recent technologies)	Advantages	Issues	Metrics used (those are used to justify the performance of the used scheme)
Title: *Evaluating deep learning architectures for Speech Emotion Recognition* ***Year**: 2020* ***Authors**:* *Haytham M. Fayek, , Margaret Lech , Lawrence Cavedon*	Deep learning architecture-Deep neural networks Convolutional neural network – Feedforward network,Long short term memory.	It automatically detects the important features without any human supervision.	It is hard for a CNN to encode the position and orientation of objects. They fail to encode the position and orientation of pronounciations. They have a hard time classifying emotions with different positions.	M1 : Correctness of the input M2: Layer segregation measures M3: Accuracy of emotions

***Title**:* *Speech emotion recognition using hidden Markov models* ***Y ear**:2021* ***Authors**:* *Tin Lay Nwe , Say Wei Foo , Liyanage C. De Silva*	Short time log frequency power coefficients and Hidden Markov models.	They are the most flexible generalization of sequence profiles. It can also perform a wide variety of operations including multiple alignment, data mining and classification, structural analysis, and pattern discovery.	HMM become very complicated when more number of emotions are found in the single audio file and when more interactions among the emotions are included.	M1: Memory ability of the network M2 : Speed of identification M3 : Emotion identification accuracy
***Title**:* *Speech Emotion Recognition Using CNN* ***Y ear**:2020* ***Authors**:* *Zhengwei Huang , Ming Dong , Qirong Mao , Yongzhao Zhan*	semi-CNN based pattern recognition	By introducing a novel objective function to train semi-CNN, we can extract affect-salient features for SER by disentangling emotions from other factors such as speakers and noise.	The Semi-CNN which uses spatio- temporal learning may be more efficient than the 1D temporal learning CNN but is found not as effective as 2D spatial learning CNN in this audio based speech emotion recognition	M1 : semi-CNN performance M2 :Spatio-temporal efficiency M3 : Amount of training data

***Title**: Evaluating deep* *learning architectures for Speech Emotion Recognition* ***Year**: 2020* ***Authors**:* *Haytham M. Fayek, , Margaret Lech , Lawrence Cavedon*	Deep learning architecture-Deep neural networks Convolutional neural network – Feedforward network,Long short term memory.	It automatically detects the important features without any human supervision.	t is hard for a CNN to encode the position and orientation of objects. They fail to encode the position and orientation of pronounciations. They have a hard time classifying emotions with different positions.	M1 : Correctness of the input M2: Layer segregation measures M3: Accuracy of emotions
Title**:* *Speech emotion recognition using hidden Markov models* ***Y ear**:2021* ***Authors: *Tin Lay Nwe , Say Wei Foo , Liyanage C. De Silva*	Short time log frequency power coefficients and Hidden Markov models.	They are the most flexible generalization of sequence profiles. It can also perform a wide variety of operations including multiple alignment, data mining and classification, structural analysis, and pattern discovery.	HMM become very complicated when more number of emotions are found in the single audio file and when more interactions among the emotions are included.	M1: Memory ability of the network M2 : Speed of identification M3 : Emotion identification accuracy

Title**:* *Speech Emotion Recognition Using CNN* ***Y ear**:2020* ***Authors: *Zhengwei Huang , Ming Dong , Qirong Mao , Yongzhao Zhan*	semi-CNN based pattern recognitio	By introducing a novel objective function to train semi-CNN, we can extract affect-salient features for SER by disentangling emotions from other factors such as speakers and noise.	The Semi-CNN which uses spatio- temporal learning may be more efficient than the 1D temporal learning CNN but is found not as effective as 2D spatial learning CNN in this audio based speech emotion recognition	M1 : semi-CNN performance M2 :Spatio-temporal efficiency M3 : Amount of training data
***Title**:* Deep Learning Techniques for Speech Emotion Recognition, from Databases to Models ***Year**: 2021* ***Authors**:* : Abbaschian, B.J.; Sierra-Sosa, D.; Elmaghraby	convolutional neural networks (CNNs), recurrent neural networks (RNNs), and hybrid models such as the convolutional recurrent neural network (CRNN)	it provides a critical analysis of the strengths and weaknesses of different deep learning models and their respective architectures and training procedures.	some of the deep learning models and techniques discussed in the review are relatively new, and their practical usefulness and effectiveness may not have been thoroughly tested in real-world scenarios.	M1: accuracy M2 : F1-score M3 : Confusion matrix M4: area under curve

Title: Clustering-Based Speech Emotion Recognition by Incorporating Learned Features and Deep BiLSTM ***Y ear****:2019* ***Authors****:* Mustaqeem, M. Sajjad and S. Kwon	pre-trained convolutional neural network (CNN	1) Generalisability 2) Effective clustering 3) Unsupervised learning 4) Incorporation of deep learning techniques	The performance of the proposed approach is dependent on the choice of hyperparameters, which may need to be tuned for different datasets and applications.	M1: accuracy M2 : F1-score M3 : Confusion matrix M4: Silhouette Coefficient
Title*:* A CNN-Assisted Enhanced Audio Signal Processing for Speech Emotion Recognition ***Year****: 2019* ***Authors****:* Mustaqeem and Soonil Kwon	audio signal processing techniques with a convolutional neural network	The proposed approach uses a CNN to extract more discriminative features from the audio signal, which can improve the performance of the emotion recognition system.	The proposed approach involves training a Convolutional Neural Network (CNN) for feature extraction, which can be computationally expensive, especially for large datasets.	M1: accuracy M2 : F1-score M3 : Confusion matrix M4: mean opinion score

Title: Two-Way Feature Extraction for Speech Emotion Recognition Using Deep Learning ***Y ear***:2022 ***Authors***: Apeksha Aggarwal , Akshat Srivastava , Ajay Agarwal , Nidhi Chahal , Dilbag Singh , Abeer Ali Alnuaim , Aseel Alhadlaq and Heung-No Lee	Decision Tree Classifier, Random Forest Classifier and a Convolutional Neural Network (CNN) Mel-Frequency Cepstral Coefficients (MFCC), Mel, Chroma, Tonnetz	Two-way feature extraction is proposed for utilizing super convergence to extract two sets of potential features from the speech data	Challenge is Scalability. Some of the feature extraction algorithms wouldn't be feasible to run if the datasets are huge. Especially the complex non-linear feature extraction methods would be infeasible.	M1: Multimodal speech data is utilized for training M2 : Amount of training data M3 : Emotion identification accuracy

Title: Robust Speech Emotion Recognition Using CNN+LSTM Based on Stochastic Fractal Search Optimization Algorithm ***Y ear***:2022 ***Authors***: Abdelaziz A. Abdelhamid , El-Sayed M. El-Kenawy , Bandar Alotaibi , Ghada M. Amer , Mahmoud Y. Abdelkader , Abdelhameed Ibrahim, Marwa Metwally Eid	deep learning model consists of a convolutional neural network (CNN) and a long short-term memory (LSTM), stochastic fractal search (SFS)-guided whale optimization algorithm	Performance monitoring can be used to track training accuracy and validation accuracy The strength of (woa) lgorithm is the ability to balance between the exploration and exploitation of the search agents' positions to guarantee to reach the optimal global solution.	Hyperparameters of deep learning models affect their performance to a certain extent. The selection of proper values of these parameters usually forms a challenge in utilizing deep learning models for different tasks.	M1:hyperparameters performance M2 :Spatio-temporal efficiency M3 : Speed of identification

***Title**:* speech emotion recognition with dual- sequence lstm architecture ***Y ear**:2020* ***Authors**:* Jianyou Wang, Michael Xue , Ryan Culhane , Enmao Diao , Jie Ding , Vahid Tarokh	Dual-Sequence LSTM ,MFCC features and mel-spectrograms	A novel mechanism for data pre-processing that uses nearest neighbour interpolation to address the problem of variable lengths between different audio signals.	While truncating and padding data, results in losing of formation and also increase the computational cost.	M1 : Correctness of the input M2: Accuracy of emotions

6. PROPOSED MODEL

Speech emotion recognition using AI is a fascinating field with numerous potential applications, ranging from improving human-computer interaction to enhancing mental health diagnostics. Here's a proposed model architecture for speech emotion recognition:

1. **Data Preprocessing**:- Audio Segmentation: Split the audio signals into smaller frames, typically ranging from 20 to 50 milliseconds, to capture temporal dynamics.- Feature Extraction: Extract relevant acoustic features from each frame, such as Mel-Frequency Cepstral Coefficients (MFCCs), Pitch, Energy, and Spectral Contrast.

2. **Feature Representation**: - Temporal Context: Incorporate contextual information by grouping multiple frames together, forming a temporal context window. This allows the model to capture temporal dependencies and variations in emotion expression over time.Feature Normalization: Normalize the extracted features to ensure consistent scaling and improve model convergence.

3. Model Architecture: - Convolutional Neural Network (CNN): Utilize 1D CNN layers to learn local acoustic patterns and capture low-level feature representations from the temporal context window. - Recurrent Neural Network (RNN): Employ bidirectional LSTM or GRU layers to model long-term dependencies and capture higher-level temporal dynamics.

- Attention Mechanism: Integrate attention mechanisms to focus on salient frames or segments in the input, enhancing the model's ability to distinguish emotional cues.

- Fusion Strategies: Combine information from different modalities, such as acoustic features, linguistic features, or facial expressions, to improve the model's

performance.

4. Training and Optimization: - Dataset: Collect a labeled dataset of speech samples annotated with emotion labels. Examples include the IEMOCAP, Emo-DB, or SAVEE datasets.

- Loss Function: Use appropriate loss functions like categorical cross-entropy or mean squared error, depending on the type of emotion classification task (e.g., discrete emotions, arousal-valence dimensions).

- Regularization Techniques: Apply regularization techniques such as dropout or batch normalization to prevent overfitting.

- Optimization: Train the model using gradient-based optimization algorithms like Adam or RMSprop, adjusting hyperparameters to achieve the best performance.

- Hyperparameter Tuning: Perform grid search or random search to find optimal hyperparameters for the model architecture and training process.

5. Evaluation and Deployment:

- Performance Metrics: Evaluate the model's performance using metrics like accuracy, precision, recall, F1-score, or area under the Receiver Operating Characteristic curve (AUC-ROC).

- Cross-Validation: Employ k-fold cross-validation to assess the model's robustness and generalization capabilities.

- Real-time Inference: Optimize the model for real-time inference on various platforms, such as mobile devices or embedded systems, considering resource constraints.

6. Further Enhancements:

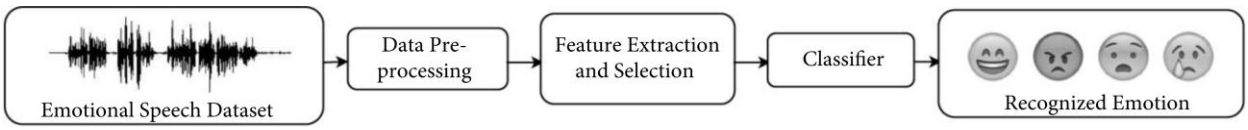
- Transfer Learning: Leverage pre-trained models on large-scale speech tasks, such as speech recognition or speaker recognition, and fine-tune them on the emotion recognition task to benefit from learned representations.

- Data Augmentation: Apply techniques like speed perturbation, noise injection, or pitch shifting to increase the diversity and size of the training dataset.

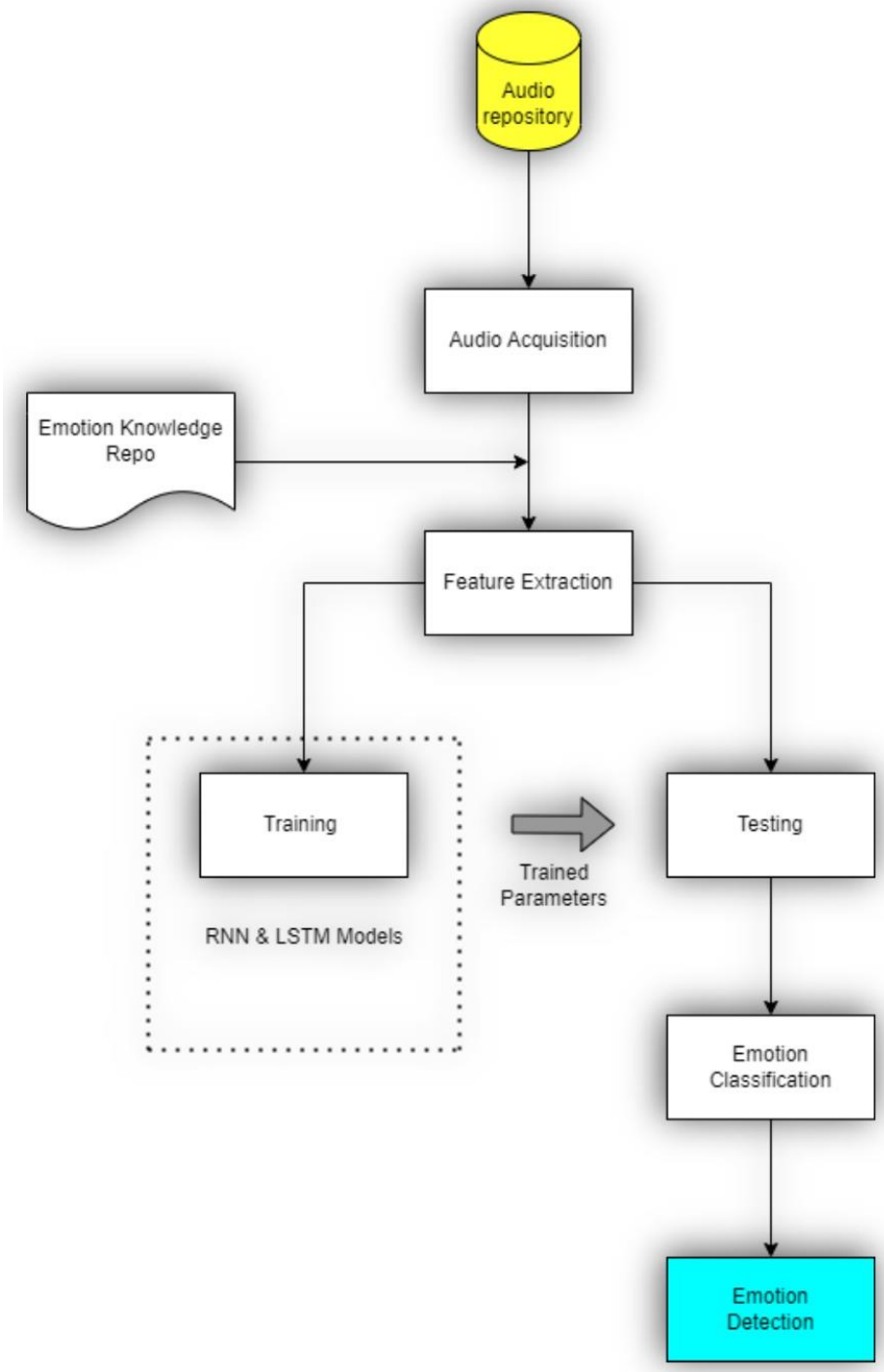
- Ensemble Methods: Combine multiple models or variations of the proposed model to improve overall performance and robustness.

Remember that the proposed model architecture serves as a general guideline, and the specific implementation may vary based on the available resources, dataset characteristics, and the desired level of complexity for the speech emotion recognition system.

7.SYSTEM ARCHITECTURE



WORKFLOW OF THE SYSTEM



7.2Detailed Description of modules

7.2.1 Importing the libraries

- **Numpy** for numerical operations
- **Pandas , Seaborn , matplotlib** for data visualisation
- **sklearn** to implement machine learning models
- **IPython** interactive command shell for python
- **Tensorflow** helps create flowgraohs and process series of nodes, import keras layers

7.2.2 Loading the dataset

The dataset chosen for this project is CREMA-D which is loaded into the input section of Kaggle IDE. And create arrays to store the emotion path and the emotion label. Create a data frame to storethe corresponding emotion to the file path.

7.2.3 Data visualisation

Using the seaboard count plot the total count of each. Emotion for the given dataset.

For the next visualisation, create a wave plot using def create_waveplot and specify the fig size.Create a spectrogram function and apply for all the emotions

7.2.4 MFCC Extraction

Create a directory named labels that maps emotions with numerical value. Set the number of mfccsto be extracted, sampling rate and the hop length from one frame to another. Create a loop which iterates through all the audio files to extract the mfcc. The loop shows iteration of the files being extracted. Add padding bits to the extracted mfcc files using keras library

7.2.5 Splitting the data

Split the data into testing and training data sets. Specify the test dataset size.

7.2.6 Model Building

Build a model with a preferred input shape. This model has 5 layers where the first LSTM

The first layer consist of 124 layers that inputs the shape to the second layer of LSTM which has 64layers , a dense layer having 64 layers , a dropout layer , and a dense layer having 6 layers.

The first dense layer has tanh as activation function and the second dense layer have softmaxactivation function for categorically classifying the output.

The optimiser used here is Adam and the learning rate of the model is 0.001 .

7.2.6 Training

The model trains by fitting the model with X_train and y_train with a batch size 64 and by tuningthe hyper parameters, we can find the accuracy of the 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:

IMPORTING LIBRARIES:

```
import pandas as pd
import numpy as np

import os
import sys

import librosa
import librosa.display
import seaborn as sns
import matplotlib.pyplot as plt

from sklearn.preprocessing import StandardScaler, OneHotEncoder
from sklearn.model_selection import train_test_split

from IPython.display import Audio

import tensorflow as tf
from tensorflow.keras.callbacks import ReduceLROnPlateau
from tensorflow.keras.models import Sequential
from tensorflow.keras.layers import Dense, LSTM, Flatten, Dropout
from tensorflow.keras.callbacks import ModelCheckpoint

import warnings
if not sys.warnoptions:
    warnings.simplefilter("ignore")
warnings.filterwarnings("ignore", category=DeprecationWarning)
```

DATASET

```
Crema = "/kaggle/input/cremad/AudioWAV/"

crema_directory_list = os.listdir(Crema)

file_emotion = []
file_path = []

for file in crema_directory_list:
    # storing file paths
    file_path.append(Crema + file)
    # storing file emotions
    part=file.split('_')
    if part[2] == 'SAD':
        file_emotion.append('sad')
    elif part[2] == 'ANG':
        file_emotion.append('angry')
    elif part[2] == 'DIS':
        file_emotion.append('disgust')
    elif part[2] == 'FEA':
        file_emotion.append('fear')
    elif part[2] == 'HAP':
        file_emotion.append('happy')
    elif part[2] == 'NEU':
        file_emotion.append('neutral')
    else:
        file_emotion.append('Unknown')

# dataframe for emotion of files
emotion_df = pd.DataFrame(file_emotion, columns=['Emotions'])

# dataframe for path of files.
```

```
path_df = pd.DataFrame(file_path, columns=['Path'])
Crema_df = pd.concat([emotion_df, path_df], axis=1)
Crema_df.head()
```

DATA VISUALISATION

```
plt.title('Count of Emotions', size=16)
sns.countplot(Crema_df.Emotions)
plt.ylabel('Count', size=12)
plt.xlabel('Emotions', size=12)
sns.despine(top=True, right=True, left=False, bottom=False)
plt.show()

def create_waveplot(data, sr, e):
    plt.figure(figsize=(10, 3))
    plt.title('Waveplot for {} emotion'.format(e), size=15)
    librosa.display.waveplot(data, sr=sr)
    plt.show()

def create_spectrogram(data, sr, e):
    X = librosa.stft(data)  #librosa.stft() converts time domain to frequency domain
    Xdb = librosa.amplitude_to_db(abs(X))
    plt.figure(figsize=(12, 3))
    plt.title('Spectrogram for {} emotion'.format(e), size=15)
    librosa.display.specshow(Xdb, sr=sr, x_axis='time', y_axis='hz')
    plt.colorbar()

emotion='angry'
path = np.array(Crema_df.Path[Crema_df.Emotions==emotion])[0]
data, sampling_rate = librosa.load(path)
create_waveplot(data, sampling_rate, emotion)
create_spectrogram(data, sampling_rate, emotion)
Audio(path)
```

MFCC EXTRACTION

```
#we are creating a directory named labels that maps emotions with a numerical value
labels = {'disgust':0,'happy':1,'sad':2,'neutral':3,'fear':4,'angry':5}
Crema_df.replace({'Emotions':labels},inplace=True) #after allocating numbers, the numbers will replace emotion


num_mfcc=13
n_fft=2048
hop_length=512
SAMPLE_RATE = 22050
data = {
        "labels": [],
        "mfcc": []
    }
#for all 7442 files, the emotion label is added. librosa.feature.mfcc extraction. i%500 iterate the progress of extraction
for i in range(7442):
    data['labels'].append(Crema_df.iloc[i,0])
    signal, sample_rate = librosa.load(Crema_df.iloc[i,1], sr=SAMPLE_RATE)


    mfcc = librosa.feature.mfcc(signal, sample_rate, n_mfcc=13, n_fft=2048, hop_length=512)
    mfcc = mfcc.T
    data["mfcc"].append(np.asarray(mfcc))
    if i%500==0:
        print(i)
```

PADDING THE MFCC BITS

```
X = np.asarray(data['mfcc'])
y = np.asarray(data["labels"])
X = tf.keras.preprocessing.sequence.pad_sequences(X)
X.shape
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.1)
X_train, X_validation, y_train, y_validation = train_test_split(X_train, y_train, test_size=0.2)
print(X_train.shape,y_train.shape,X_validation.shape,y_validation.shape,X_test.shape,y_test.shape)
```

BUILDING THE MODEL

```
def build_model(input_shape):
    model = tf.keras.Sequential()

    model.add(LSTM(124, input_shape=input_shape, return_sequences=True))
    model.add(LSTM(64))

    model.add(Dense(64, activation='tanh'))
    model.add(Dropout(0.3)) #0.33

    model.add(Dense(6, activation='softmax'))

    return model
# create network
input_shape = (None,13)
model = build_model(input_shape)
# compile model
optimiser = tf.keras.optimizers.Adam(learning_rate=0.001) #0.001
model.compile(optimizer=optimiser,
              loss='sparse_categorical_crossentropy',
              metrics=['accuracy'])

model.summary()
```

TRAINING

history = model.fit(X_train, y_train, validation_data=(X_validation, y_validation), batch_size=64, epochs=50)

EVALUATION

test_loss, test_acc = model.evaluate(X_test, y_test, verbose=0)

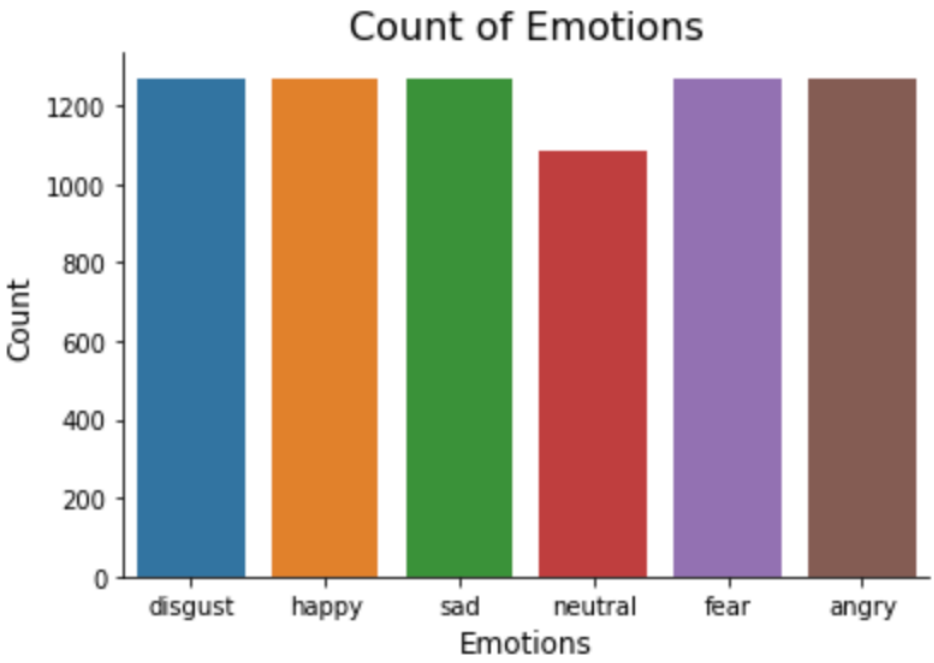
print("Test Accuracy: ",test_acc)

9.2 SCREENSHOTS OF OUTPUTS:

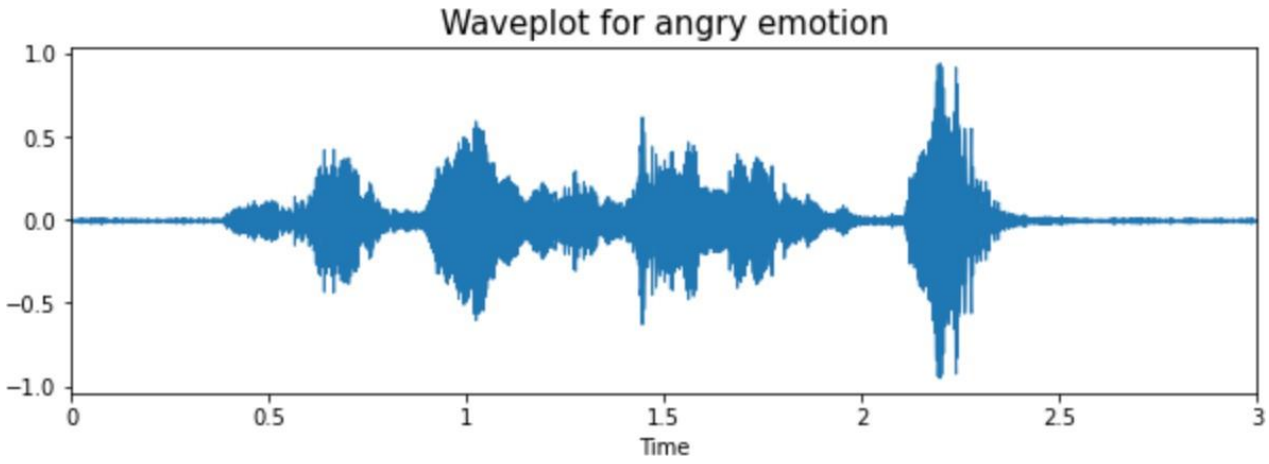
Dataframe of emotion file paths: after splitting the data path to acquire the emotion label,we retrieve it and put eat data path to its connected emotion

	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

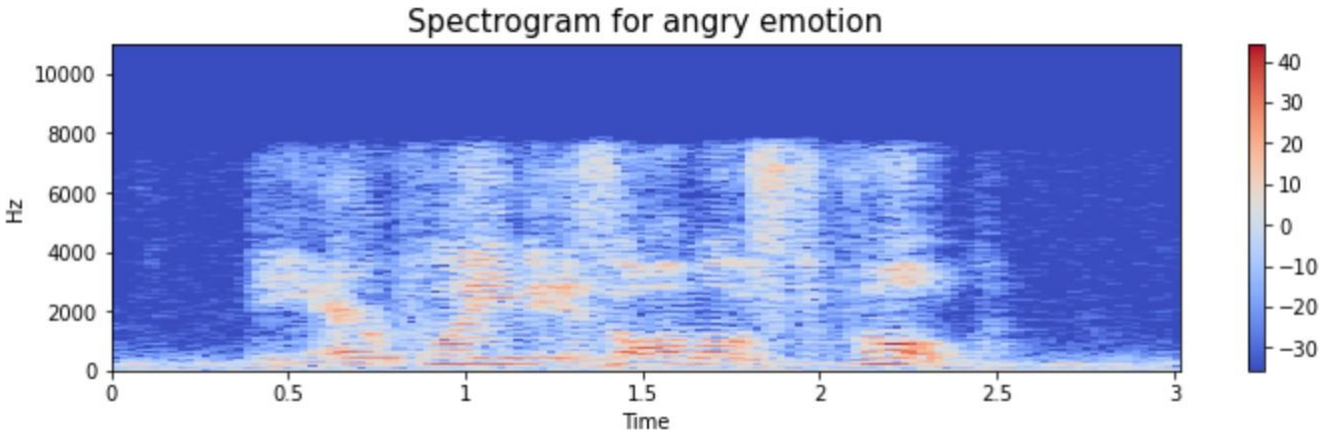
Dataplot of emotions: no of files for each emotion is counted using the seaborn library in python



Wave plot: plot between the amplitude (y-axis) and time (x-axis) for a given audio. It maps thesampling rate of the given audio. It captures the frames of audio for given time unit



Spectrogram plots the graph between frequency (y-axis) and time(x-axis) for the given emotionaudio.



Model Summary: the summary gives the number of layers and the types of layer the said model isbuilt on. The first layer being LSTM with 124 layers of neural network has been built first and the second LSTM layer has 64 layers of neutrons. The third layer is dense with 64 layers and there is adropout layer which has 0.3% of data to be dropped out. This is essential because it avoids overfitting and helps the machine learn better.The final layer is the dense layer which consolidates into 6 layers each belonging to the 6 emotions namely - disgust, sad, happy, angry, neutral, fear .

```
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
```

Running the model:

Using the model.fit() command, the model runs on X_train and y_train values . The data split into training and validation comes into play in this part where the training data is fed to the model with adefined batch size i.e., 64. Each time for each epoch, batches of 64 will be jumbled and sent to the model to learn better. The assurance used here is the accuracy metrics to ensure that model works togive better accuracy.

```
history = model.fit(X_train, y_train, validation_data=(X_validation, y_validation), batch_size=64, epochs=50)

Epoch 1/50
84/84 [==============================] - 46s 514ms/step - loss: 1.6899 - accuracy: 0.2552 - val_loss: 1.4443 - val_accuracy: 0.4022
Epoch 2/50
84/84 [==============================] - 42s 497ms/step - loss: 1.4842 - accuracy: 0.3860 - val_loss: 1.4265 - val_accuracy: 0.4351
Epoch 3/50
84/84 [==============================] - 42s 498ms/step - loss: 1.4398 - accuracy: 0.4002 - val_loss: 1.3565 - val_accuracy: 0.4530
Epoch 4/50
84/84 [==============================] - 42s 505ms/step - loss: 1.3696 - accuracy: 0.4476 - val_loss: 1.3809 - val_accuracy: 0.4522
Epoch 5/50
84/84 [==============================] - 42s 497ms/step - loss: 1.3507 - accuracy: 0.4627 - val_loss: 1.3206 - val_accuracy: 0.4784
Epoch 6/50
84/84 [==============================] - 42s 495ms/step - loss: 1.3339 - accuracy: 0.4691 - val_loss: 1.3578 - val_accuracy: 0.4463
Epoch 7/50
84/84 [==============================] - 41s 493ms/step - loss: 1.3296 - accuracy: 0.4677 - val_loss: 1.3086 - val_accuracy: 0.4836
Epoch 8/50
84/84 [==============================] - 41s 493ms/step - loss: 1.2921 - accuracy: 0.4758 - val_loss: 1.3142 - val_accuracy: 0.4761
Epoch 9/50
84/84 [==============================] - 41s 486ms/step - loss: 1.2662 - accuracy: 0.4962 - val_loss: 1.2920 - val_accuracy: 0.4806
Epoch 10/50
84/84 [==============================] - 42s 501ms/step - loss: 1.2428 - accuracy: 0.5105 - val_loss: 1.2831 - val_accuracy: 0.4866
Epoch 11/50
84/84 [==============================] - 42s 498ms/step - loss: 1.2428 - accuracy: 0.5065 - val_loss: 1.2319 - val_accuracy: 0.5075
Epoch 12/50
84/84 [==============================] - 42s 502ms/step - loss: 1.2066 - accuracy: 0.5212 - val_loss: 1.2358 - val_accuracy: 0.5082
Epoch 13/50
84/84 [==============================] - 42s 502ms/step - loss: 1.1665 - accuracy: 0.5444 - val_loss: 1.2445 - val_accuracy: 0.5201
Epoch 14/50
84/84 [==============================] - 42s 495ms/step - loss: 1.1815 - accuracy: 0.5373 - val_loss: 1.1728 - val_accuracy: 0.5284
Epoch 15/50
84/84 [==============================] - 41s 489ms/step - loss: 1.1642 - accuracy: 0.5444 - val_loss: 1.1905 - val_accuracy: 0.5187
```

Evaluation

```
[ ] test_loss, test_acc = model.evaluate(X_test, y_test, verbose=0)
    print("Test Accuracy: ",testacc)

    Test Accuracy:  0.5699999928474426

[ ] model.save('Speech-Emotion-Recognition-Model.h5')

    def predict_emotion(file_name):
        mfccs = extract_features(file_name)
        if mfccs is not None:
            mfccs = np.array([mfccs])
            mfccs = np.transpose(mfccs, (0, 2, 1))
            prediction = model.predict(mfccs)[0]
            prediction_label = le.inverse_transform([np.argmax(prediction)])
            return prediction_label[0]
        else:
            rand_emo()

[ ] file_name = "/kaggle/input/emynvomit"
    predict_emotion(file_name)
    print("is the Predicted Emotion")

    0
    3000
    6000
    Fear
    is the Predicted Emotion
```

10.REFERENCES

1) https://doi.org/10.3390/s21041249

2) https://doi.org/10.3390/s20010183

3) https://www.researchgate.net/publication / 340946514_Attention-LSTM-Attention_Model_for_Speech_Emotion_Recognition_and_Analysis_of_IEMOCAP_Da tabase

4) https://www.ijraset.com/research-paper/speech-emotion-trecognition- system-using-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/9078789

9) Nogueiras, A., Moreno, A., Bonafonte,

10) A. and Marino, J., B.,"Speech emotion recognition using hidden Markov models." *INTERSPEECH*, pp. 2679–2682, 2001.

11) Busso, C., Lee, S., Narayanan, S., "Analysis of emotionally salient aspects of fundamental frequency for emotion detection," *IEEE Trans- actions on Audio, Speech and Language Processing,* vol.17, no. 4, pp. 582–596, 2009.

12) El Ayadi, M., Kamel, M., S., Karray, F., "Survey on speech emotion recognition: Features, classification schemes, and databases," *Pattern Recognition,* vol. 44, no. 3, pp. 572–587, 2011.

13) Shashidhar, G., K., and Sreenivasa, K., R., "Emotion recognition from speech: a review," *International journal of speech technology*, vol. 15, no. 2, pp. 99–117, 2012

14) Mower, E., Mataric, M., J. and Narayanan, S., "A framework for automatic human emotion classification using emotion profiles," *IEEE Transactions on Audio, Speech, and Language Processing,* vol. 19, no. 5, pp. 1057–1070, 2011.

15) Schuller, B., Steidl, S., Batliner, A., Burkhardt, F., Devillers, L., Müller, C., Narayanan, S., "The INTERSPEECH 2010 Paralinguistic Challenge," *INTERSPEECH,* pp. 2794–2797, 2010

16) Chorowski, J., K., Bahdanau, D., Serdyuk, D., Cho, K. and Bengio, Y., "Attention-based models for speech recognition," *Advances in Neural Information Processing Systems*, pp. 577–585, 2015.

17) Huang, C., W., Narayanan, S., "Attention Assisted Discovery of Sub- Utterance Structure in Speech Emotion Recognition," *INTERSPEECH,* pp. 1387–1391, 2016