A FIELD PROJECT REPORT

On

“SPEECH EMOTION RECOGNITION”

Submitted

by

221FA04112	221FA04382
Sathvika	Yojitha
221FA04435	221FA04437
Niharika	Nithya Sri

Under the guidance of

Sajida Sultana.Sk

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING

VIGNAN'S FOUNDATION FOR SCIENCE, TECHNOLOGY AND RESEARCH Deemed to be UNIVERSITY

Vadlamudi, Guntur.

ANDHRA PRADESH, INDIA, PIN-522213.

VIGNAN'S

FOUNDATION FOR SCIENCE, TECHNOLOGY & RESEARCH

(Deemed to be University) - Estd. u/s 3 of UGC Act 1956

CERTIFICATE

This is to certify that the Field Project entitled **"Speech Emotion Recognition"** that is being submitted by 221FA04112 (Sathvika), 221FA04382(Yojitha), 221FA04435(Niharika)**,** 221FA04437(Nithya Sri) for partial fulfilment of Field Project is a bonafide work carried out under the supervision of
Sajida Sultana.Sk,   Assistant Professor, Department of CSE.

VIGNAN'S
FOUNDATION FOR SCIENCE, TECHNOLOGY & RESEARCH
(Deemed to be University) - Estd. u/s 3 of UGC Act 1956

DECLARATION

We hereby declare that the Field Project entitled **"Speech Emotion Recognition"** is being submitted by 221FA04112 (Sathvika), 221FA04382(Yojitha), 221FA04435(Niharika) and 221FA04437(Nithya Sri) in partial fulfilment of Field Project course work. This is our original work, and this project has not formed the basis for the award of any degree. We have worked under the supervision of Sajida Sultana.Sk, ,Assistant Professor, Department of CSE.

By
221FA04112 (Sathvika),
221FA04382(Yojitha),
221FA04435 (Niharika),
221FA04437 (Nithya Sri)

Date:

ABSTRACT

Speech is a commonly used signal for interaction between humans, this leads to the usage of speech for human and machine interactions as well. Improvements in this interactive system reach toward speech emotion recognition (SER) system.SER gives sufficient intelligence for efficient natural communication between humans and machines. SER system classifies emotional states such as sadness, angry, neutral, and happiness from the speaker's utterances. This paper describes speech features and machine learning models that can be used for SER. For effective classification and to learn multidimensional complex data, a deep learning algorithm is used in this system. This paper also presents the preliminary results of a system with an MFCC feature and an LSTM algorithm. To enhance the system's performance, data augmentation techniques and feature extraction methods like spectral contrast and chroma features are considered. The proposed system's architecture is optimized through hyperparameter tuning, which improves the classification accuracy. The results demonstrate the efficacy of combining MFCC with LSTM for accurate and robust emotion detection across varying speech patterns.

TABLE OF CONTENTS

LIST OF FIGURES

CHAPTER-1

INTRODUCTION

1. INTRODUCTION

Human commonly uses a vocal language for communication. This vocal languagemotivates researchers to think for speech communication with a machine. Multiple machines are developed based on this topic like assistance applicationsin a smartphone, speech to text converter and voice command operated machines. But this system lags in natural communication with humans; this activity can be improved by giving some intelligence to a machine. A machine can understand humans more efficiently when it can recognize human perception. Speech emotion recognition (SER) helps a machine to identify human emotions and reactaccordingly.

Humans naturally convey a wealth of emotional information through their speech patterns, including variations in pitch, intonation, rhythm, and other acoustic features. SER aims to capture and interpretthese subtle cues to infer the underlying emotional state of the speaker. By analyzing speech signals, SER systems can categorize emotions into discrete categories such as happiness, sadness, anger, fear, surprise, disgust, or neutrality.

1.1 What is Speech Emotion Recognition?

Speech Emotion Recognition (SER) is the process of automatically identifying the emotional state of a speaker based on their speech signal. It involves analyzing various acoustic features present in the speech signal to infer the speaker's emotional state accurately. These emotional states can include happiness, sadness, anger, fear, surprise, disgust, or neutrality.

1.2 Importance of SER

Speech Emotion Recognition (SER) is pivotal due to its multifaceted contributions across various domains:

1. Enhanced Human-Computer Interaction: SER empowers machines to comprehend human emotions, fostering more intuitive interactions. This advancement is vital for applications like virtual assistants and customer service bots, enhancing user experience and satisfaction.

2. Insight into Emotional States: Understanding emotions is crucial for deciphering human behavior and decision-making processes. SER provides valuable insights into emotional states, benefiting fields such as psychology, market research, and social sciences.

3. Personalization and Adaptation: By discerning emotional cues, SER enables tailored experiences. This personalization extends to recommendation systems, educational software, and adaptive learning platforms, enriching user engagement and learning outcomes.

4. Mental Health Monitoring: SER aids in the early detection and management of mental health disorders by analyzing speech patterns indicative of conditions like depression and anxiety. This contributes to personalized healthcare and early intervention strategies.

5. Improved Safety and Security: Integrating SER into security systems enables the detection of distress signals or abnormal emotional states in emergency situations. For instance, SER technology can assist in identifying callers in distress in call center environments, prioritizing urgent assistance.

6. Market Research and Customer Insight: SER facilitates the analysis of customer feedback and sentiment, guiding businesses in making informed decisions. By understanding customer emotions, companies can refine products, services, and marketing strategies to enhance customer satisfaction and loyalty.

7. Entertainment and Gaming: In the realm of entertainment, SER enhances gaming experiences by adapting game dynamics based on players' emotional responses. This real-time adaptation creates more immersive and engaging gaming experiences.

1.3 Applications of SER

Speech Emotion Recognition (SER) finds diverse applications across numerous domains:

1. Customer Service: SER assists in analyzing customer interactions to gauge satisfaction levels, detect frustration, and enhance service quality in call centers and customer support services.

2. Healthcare: SER aids in mental health monitoring by detecting speech patterns indicative of mood disorders, facilitating early intervention and personalized treatment plans.

3. Education: SER enhances educational software and adaptive learning systems by tailoring content and activities based on students' emotional responses, improving engagement and learning outcomes.

4. Human-Computer Interaction: SER enables more natural and empathetic interactions with virtual assistants, chatbots, and other intelligent systems, enhancing user experience and engagement.

5. Market Research: SER contributes to market research by analyzing customer feedback and sentiment, guiding businesses in product development, marketing strategies, and customer satisfaction initiatives.

6. Entertainment: SER enhances gaming and interactive media experiences by adapting content and game dynamics based on players' emotional responses, creating more immersive and emotionally resonant experiences.

7. Security and Safety: SER can be integrated into security systems to detect distress signals or abnormal emotional states in emergency situations, enhancing safety and security measures.

1.4 Objectives of the Paper

The objective of the paper is to develop a Speech Emotion Recognition (SER) system using machine learning techniques, specifically focusing on the utilization of Mel- Frequency Cepstral Coefficients (MFCC) and Long Short-Term Memory (LSTM) algorithm. The paper aims to address the importance of SER in improving human-machineinteraction by enabling machines to recognize and respond to human emotions conveyed through speech. It discusses the challenges associated with speech emotion recognition, such as variations in speech due to speakers, speaking styles, and environmental factors. Additionally, the paper provides a literature survey highlighting existing methods and research in the field of SER.

2. LITERATURE SURVEY

2.1 Literature review

A literature survey is a systematic examination of existing research on a particular topic. It serves as the foundation for any scholarly investigation, offering insights into current knowledge, identifying research gaps, and providing context for new studies. By synthesizing and summarizing relevant literature, researchers can formulate precise research questions, build upon existing work, and avoid duplication. In essence, a literature survey is an essential tool for ensuring the validity and relevance of new research within the broader academic landscape.

- Pavol Harar presented a method that achieved 96.97% accuracy on testing and 69.55% on file prediction. In this method, Deep Neural Network (DNN) architecture with convolutional, pooling and fully connected layers was used for emotion recognition.
- Supriya B. Jagtap, Presented a system to detect seven emotions that are happiness, Anger, Boredom, Sadness, Surprise, Fear, and Neutral emotions. The study presents frequency information contained in speech signal are reduced into small numbers of coefficients using MFCC. This study also presents that the accuracy of the system depends on the database used for training.
- M. S. Likitha presented a method of speech emotion recognition that has proven to be 80% efficient even in a noisy environment. This system uses the MFCC feature and standard deviation values to detect emotions.
- Shumin presented methods based on LSTM-RNN models. This method has achieved 96.67% accuracy in case of angry emotion, 100% accuracy in case of sad emotion and for natural 86.67% accuracy is achieved. A literature survey shows that the MFCC feature is the most popular choice to identify emotions.

2.2 Motivation

Speech Emotion Recognition (SER) is vital for natural human-computer interaction, allowing machines to understand and respond appropriately to human emotions conveyed through speech. SER finds applications in diverse fields like affective computing, healthcare, psychology, and entertainment. It enhances user experiences in human-computer interaction by personalizing responses based on emotional cues. Moreover, SER supports assistive technologies, aids psychological research, and drives innovation in healthcare by monitoring patients' emotional well-being. Additionally, in entertainment, SER enhances gaming experiences by adapting content to users' emotional states. Overall, SER's significance lies in its ability to enable technology to empathize and connect with humans on a deeper emotional level.

3. PROPOSED SYSTEM

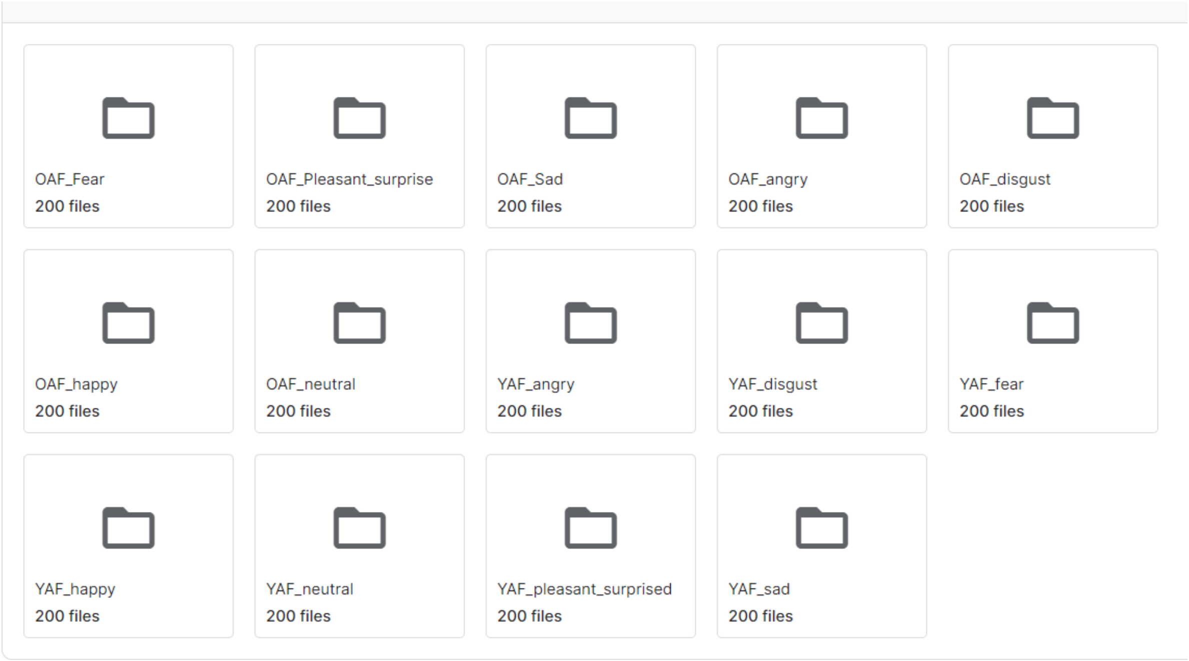
The proposed system for speech emotion recognition (SER) aims to develop an intelligent model capable of accurately detecting and classifying emotions expressed in human speech. The system leverages machine learning techniques and signal processing methodologies to analyze audio recordings and infer the underlying emotional states of the speakers. Key components of the proposed system include data collection and preprocessing, feature extraction using techniques such as Mel-Frequency Cepstral Coefficients (MFCC), selection of appropriate machine learning algorithms, model training, and evaluation. The dataset used for training and testing consists of diverse audio samples annotated with corresponding emotional labels.During preprocessing, techniques like noise removal, normalization, and feature scaling are applied to enhance the quality of the audio data. Feature extraction extracts relevant acoustic features from the audio signals, which serve as input to the machine learning models. Various machine learning algorithms, such as deep learning models like Long Short-Term Memory (LSTM) networks or traditional classifiers like Support Vector Machines (SVM), are evaluated for their effectiveness in emotion recognition tasks. The performance of the SER model is assessed using metrics like accuracy, precision, recall, and F1-score.

Overall, the proposed system aims to provide an efficient and reliable solution for automatically recognizing emotions from speech signals, with potential applications in areas such as human-computer interaction, sentiment analysis, and affective computing.

3.1 Input dataset

The dataset used in the presented system is Toronto emotional speech set (TESS). There are a set of 200 target words were spoken in the carrier phrase "Say the word _' by two actresses (aged 26 and 64 years) and recordings were made of the set portraying each of seven emotions (anger, disgust, fear, happiness, pleasant surprise, sadness, and neutral). There are 2800 data points (audio files) in total.

The dataset is organised such that each of the two female actor and their emotions are contain within its own folder. And within that, all 200 target words audio file can be found. The format of the audio file is a WAV format.



OAF_Fear
200 files

OAF_Pleasant_surprise
200 files

OAF_Sad
200 files

OAF_angry
200 files

OAF_disgust
200 files

OAF_happy
200 files

OAF_neutral
200 files

YAF_angry
200 files

YAF_disgust
200 files

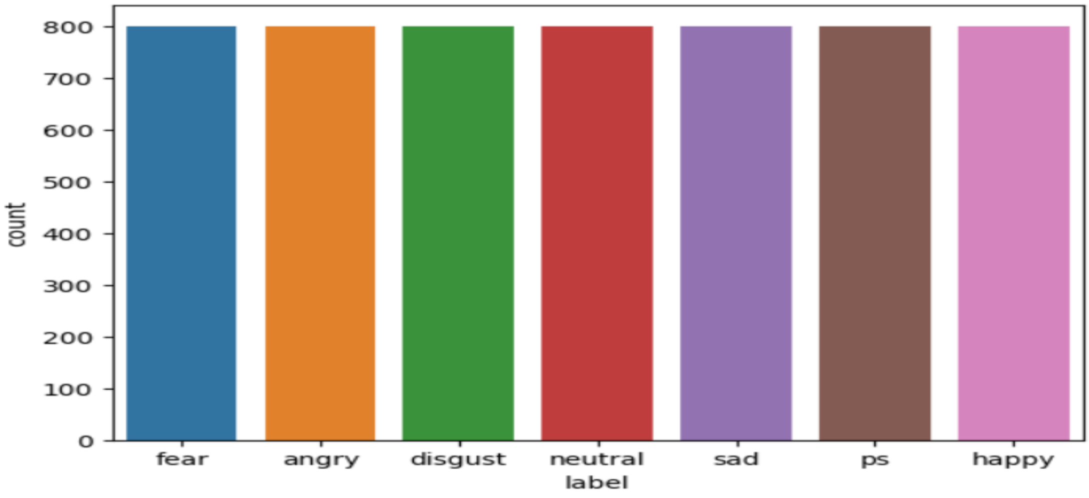
YAF_fear
200 files

YAF_happy
200 files

YAF_neutral
200 files

YAF_pleasant_surprised
200 files

YAF_sad
200 files



3.2 Data Pre-processing

Data pre-processing is the essential process of preparing raw data for analysis and modelling by cleaning, transforming, and structuring it to enhance data quality and utility. It involves tasks like handling missing values, correcting errors, encoding features, and scaling data to ensure it's in an optimal form for further analysis. It encompasses a range of operations and transformations designed to refine raw data, ensuring that it is clean, structured, and amenity subsequent analysis. This process is driven by its manifold significance in data science and analysis.

Through meticulous data cleaning, transformation, feature engineering, dimensionality reduction, outlier handling, scaling, and data splitting, it prepares raw data for more accurate and reliable analysis and modelling. Ultimately, the goal is to obtain more meaningful insights, make informed decisions, and optimize predictive models for a wide range of applications in data science and analysis.

Here are some common techniques used for pre-processing sound data in the context of sound emotion recognition (SER):

1. Noise Reduction:

- **Goal:** Eliminate unwanted background noise that can interfere with emotion-related features.
- **Techniques:**
  - **Filtering:** Applying filters to remove specific frequencies associated with the noise, like power line hum or traffic rumble.
  - **Spectral subtraction:** Estimating and subtracting the noise spectrum from the original signal.
  - **Statistical methods:** Utilizing statistical properties of the noise and speech to separate them.

2. Silence Removal:

- **Goal:** Remove silent segments irrelevant to emotion recognition, improving efficiency and avoiding feeding irrelevant information to the model.
- **Techniques:**
  - **Energy-based detection:** Identifying and removing segments with energy levels below a predefined threshold.

    - **Spectral features:** Utilizing spectral properties like zero-crossing rate or spectral entropy to detect silence.

3. Feature Scaling:

- **Goal:** Ensure all features are within a similar range, preventing one feature from dominating the model's learning.
- **Techniques:**
    - **Normalization:** Rescaling features to a range between 0 and 1.
    - **Standardization:** Subtracting the mean and dividing by the standard deviation of each feature.

4. Data Augmentation (Optional):

- **Goal:** Artificially increase the size and diversity of the training data to prevent overfitting and improve model generalization.
- **Techniques:**
    - **Adding noise:** Introducing controlled levels of background noise to simulate real-world scenarios.
    - **Speed perturbation:** Slightly increasing or decreasing the playback speed of the audio to create variations in emotional expression.
    - **Pitch shifting:** Adjusting the fundamental frequency of the audio to simulate different speaker characteristics.

5. Missing Value Handling:

- **Goal:** Address missing values in the audio data that can occur due to various reasons.
- **Techniques:**
    - **Interpolation:** Estimating missing values based on surrounding data points.
    - **Deletion:** Removing rows or columns with a high percentage of missing values, especially if they don't significantly impact the overall data.
    - **Mean/median imputation:** Replacing missing values with the average or median value of the feature.

6. Data Windowing:

- **Goal:** Divide the audio signal into smaller segments (windows) to capture the temporal dynamics of emotion within an utterance.
- **Techniques:**
    - **Rectangular window:** Simplest approach, but can introduce abrupt discontinuities at the window edges.

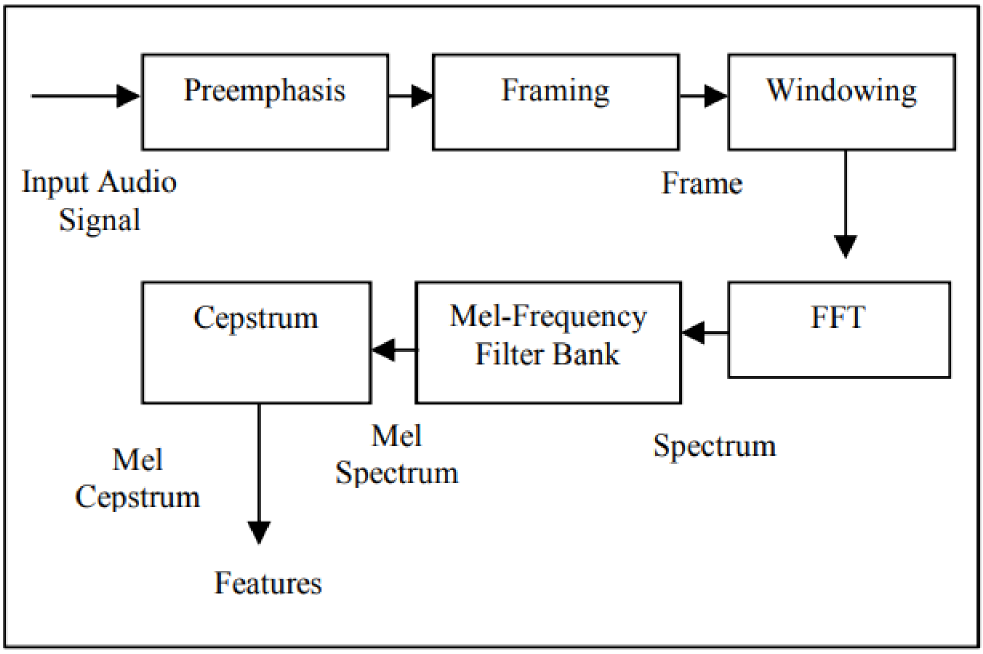
- **Hanning window:** Commonly used, smoothens the edges of the window to reduce spectral leakage.

7. Framing:

- **Goal:** Extract short-term features from each windowed segment.
- **Techniques:**
    - Overlapping windows: Ensures continuity between adjacent frames and captures information across window boundaries.

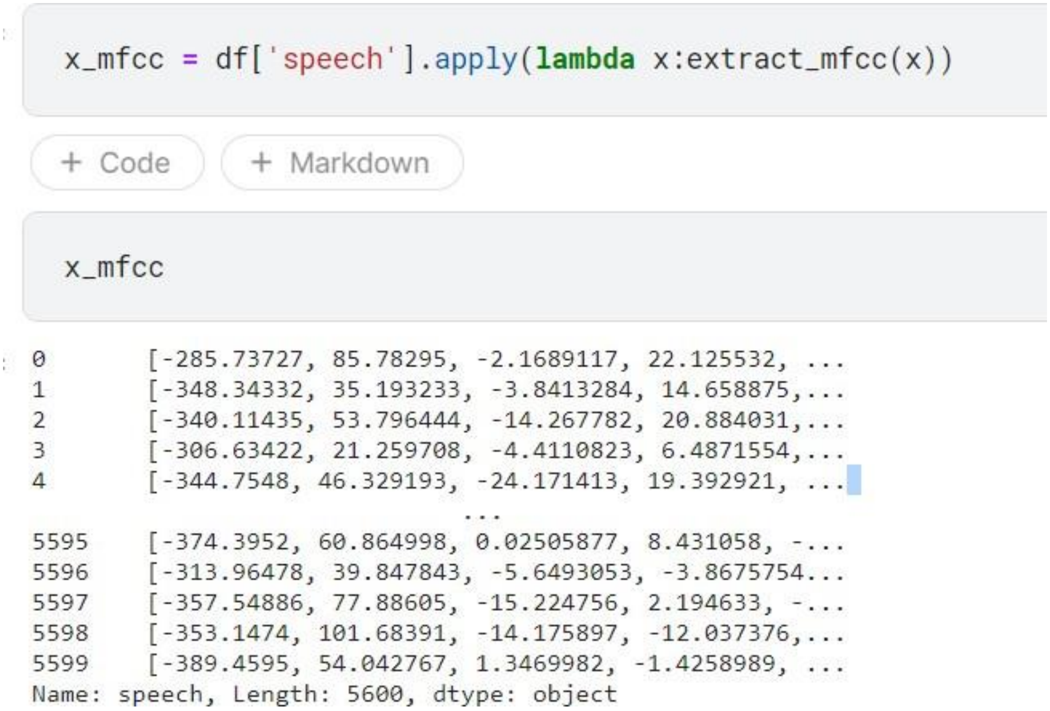
3.3 Feature extraction

MFCC provides a high level of perception of the human voice and achieving high recognition accuracy . MelFrequency Cepstral Coefficient (MFCC) is a popular and powerful analytical tool in the field of speech recognition. MFCC reduces the computational complexity of the approach, gives better ability to extract the features and can find the different parameters like pitch and energy. Mel Frequency Cepstral Coefficient (MFCC) reduces the frequency information of speech signal into the small number of coefficients which is easy to compute and extract the features. It represents the short-term power spectrum of sound, based on linear cosine transform of a log power spectrum on a non-linear Mel scale of frequency. A survey shows that the MFCC gives good results as compared to other features for a speech-based emotion recognition system .



Block diagram of MFCC process

Above figure shows the block diagram of the MFCC process. Pre-emphasis is the process of the speech signal in which pre-emphasis filter is used to achieve a smoother spectral. In the frame blocking process, the sound signal is segmented into multiple small overlapped frames in the presented methodology frame size of 20ms and the step between successive frames is also 20ms. Windowing is a process required for analyzing a section of long signals. This process removes the aliasing. Fast Fourier Transform (FFT) is used to convert a time-domain signal into a frequency spectrum. Mel-Frequency filter bank used for converting a linear frequency scale to the Mel-frequency scale. Mel-frequency scale is designed according to the perception of the human ear against the sound frequency. The scale of MelFrequency is a logarithmic scale, so it is sensitive to a lower frequency than a higher frequency. In the cepstrum process, Mel- spectrum will be converted into the time domain by using a Discrete Cosine Transform (DCT) to get the Melfrequency Cepstrum coefficient (MFCC).



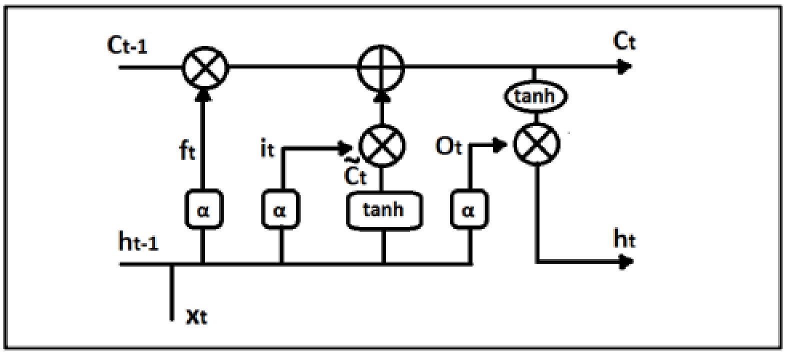
```
x_mfcc = df['speech'].apply(lambda x:extract_mfcc(x))
```

```
x_mfcc
```

```
0       [-285.73727, 85.78295, -2.1689117, 22.125532, ...
1       [-348.34332, 35.193233, -3.8413284, 14.658875,...
2       [-340.11435, 53.796444, -14.267782, 20.884031,...
3       [-306.63422, 21.259708, -4.4110823, 6.4871554,...
4       [-344.7548, 46.329193, -24.171413, 19.392921, ...
                                ...
5595    [-374.3952, 60.864998, 0.02505877, 8.431058, -...
5596    [-313.96478, 39.847843, -5.6493053, -3.8675754...
5597    [-357.54886, 77.88605, -15.224756, 2.194633, -...
5598    [-353.1474, 101.68391, -14.175897, -12.037376,...
5599    [-389.4595, 54.042767, 1.3469982, -1.4258989, ...
Name: speech, Length: 5600, dtype: object
```

3.4 Model building

ML model used in the presented methodology is based on LSTM architecture. Long short-term memory (LSTM) is a modified version of artificial recurrent neural network (RNN) architecture. LSTM works better with a huge amount of data and enough training data. The main advantage of RNN over ANN is in the case of a sequence of data it gives better performance. In the case of speech processing signal is framed in small pieces this small section, for emotion detection the dependency of each section with the previous one should be considered. So in this case LSTM gives better performance.



LSTM STRUCTURE

$$i_t = \sigma(x_t U^i + h_{t-1} W^i) \tag{1}$$

$$f_t = \sigma(x_t U^f + h_{t-1} W^f) \tag{2}$$

$$O_t = \sigma(x_t U^o + h_{t-1} W^o) \tag{3}$$

$$\tilde{C}_t = \tanh(x_t U^g + h_{t-1} W^g) \tag{4}$$

$$C_t = \sigma\left(f_t * C_{t-1} + i_t * \tilde{C}_t\right) \tag{5}$$

$$h_t = \tanh(C_t) * O_t \tag{6}$$

The above figure shows the structure of the LSTM cell and equations describing the LSTM model. U is the weight matrix contains the inputs to the hidden layer, W is the connection between current and previous layer. C is the internal memory of the unit, which is a combination of the previous memory, multiplied by the for-get gate, and the newly computed hidden state, multiplied by the input gate. ct (bar) is a candidate hidden state that is computed based on the current input and the previous hidden state .

3.5 Methodology of the system

Having discussed the foundational elements in the preceding sections, we now venture into the core of our traffic congestion prediction system. In this section, we embark on a journey through the inner workings of our model, unveiling the methodology that drives our system's ability to forecast traffic congestion. Just as a well-orchestrated symphony requires each instrument to play its part harmoniously, our methodology combines data, pre-processing, modelling, and evaluation to create a seamless and efficient prediction system.

The proposed speech emotion recognition (SER) system utilizes a combination of Mel-Frequency Cepstral Coefficients (MFCC) features and a Long Short-Term Memory (LSTM) network architecture. Here's a breakdown of the likely architectural components:

1. Pre-processing:

- Input: Raw audio signal.

- Steps:
    - May involve noise reduction, silence removal, and other techniques to improve data quality.
    - Not explicitly mentioned in the abstract, but likely present in a complete system.

2. Feature Extraction:

- Input: Pre-processed audio signal.
- Method: MFCC extraction:
    - Segmenting the audio into short frames.
    - Applying windowing functions to smooth the signal.
    - Calculating the Mel-frequency cepstrum, a representation of the short-term power spectrum of the audio on a Mel scale (perceptually relevant scale for human hearing).

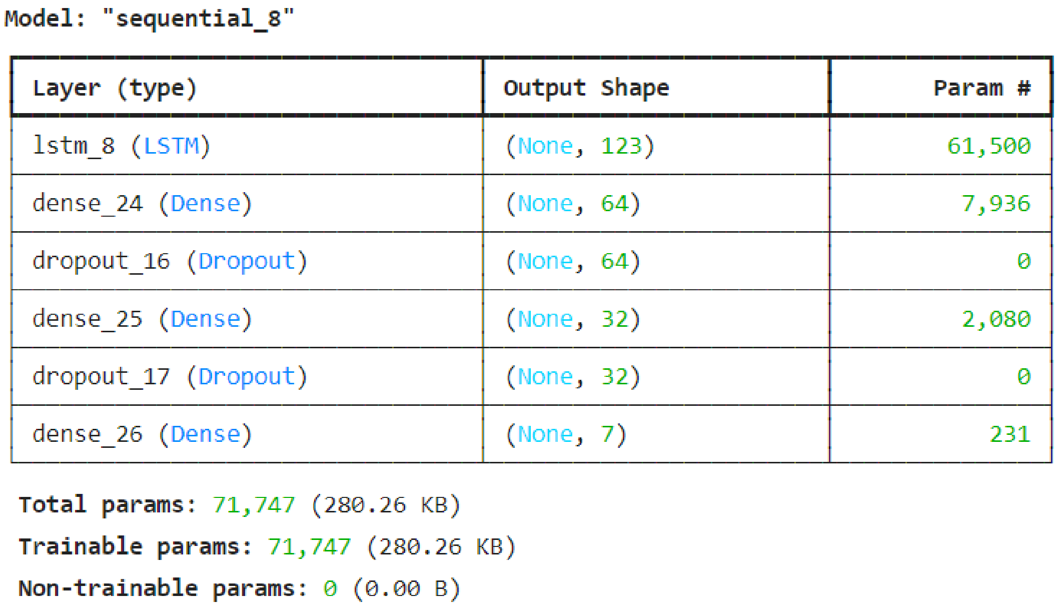
3. Machine Learning Model:

- Input: Extracted MFCC features.
- Architecture: LSTM network:
    - A type of recurrent neural network (RNN) capable of capturing long-term dependencies in sequential data like speech.
    - Likely structure based on the abstract:
        - Input layer: Receives the extracted MFCC features.
        - Hidden layers: One or more LSTM layers responsible for learning temporal relationships in the features.
        - Output layer: Produces predictions for the emotions present in the speech (e.g., happy, angry, sad, etc.).

4. Output:

- The model outputs the predicted emotions based on the processed audio signal.

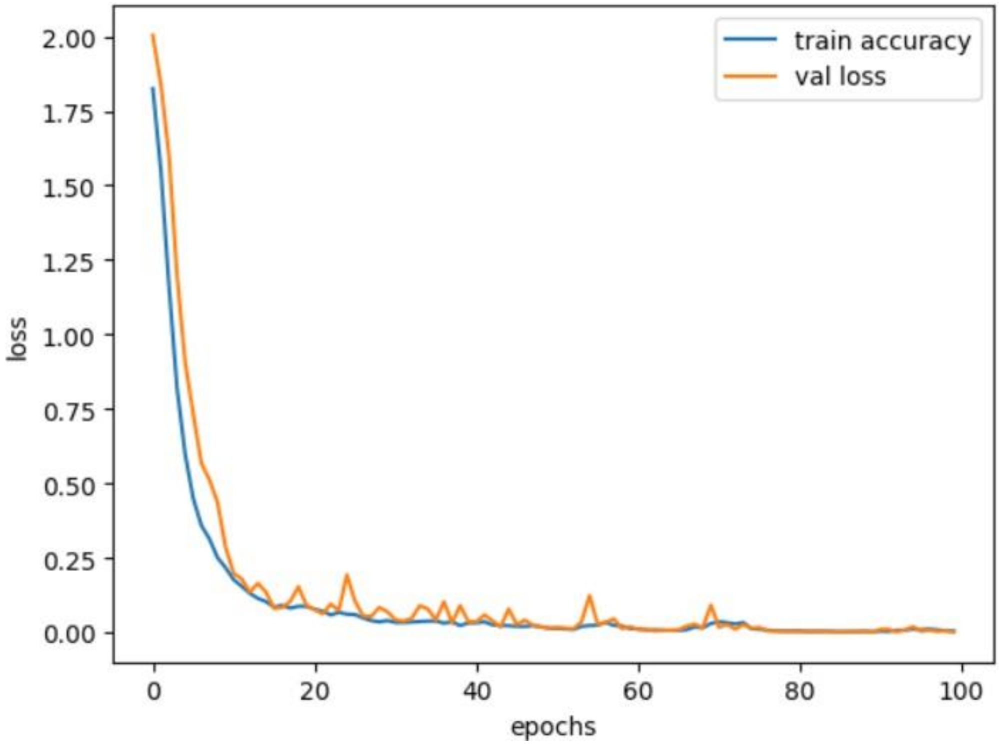
3.6 Model evaluation

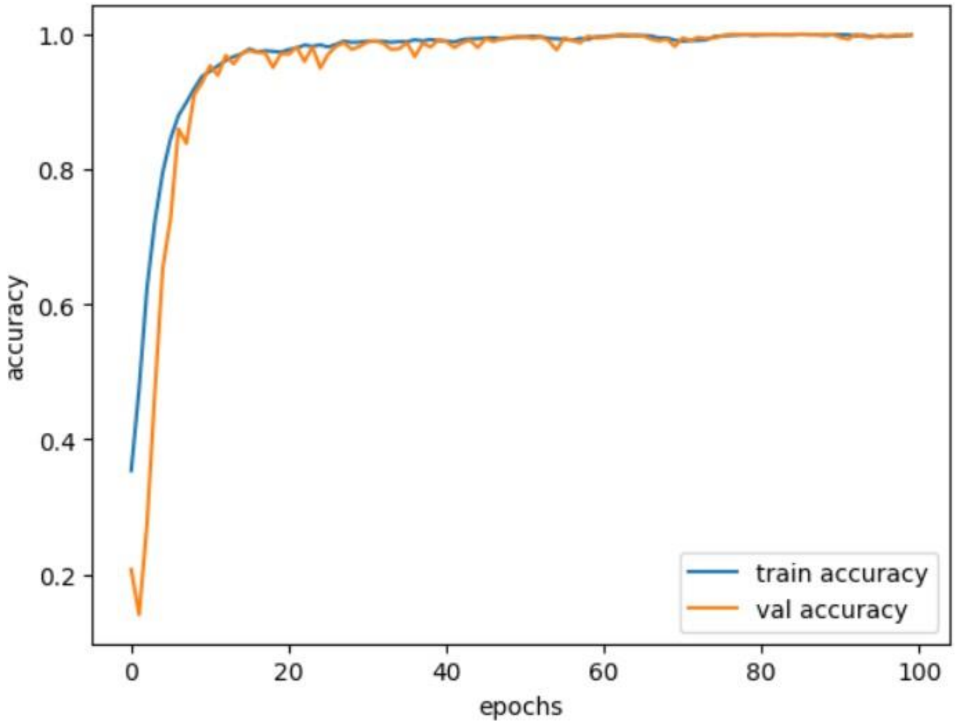


```
Model: "sequential_8"

┌──────────────────────────────┬──────────────────────┬─────────────┐
│ Layer (type)                 │ Output Shape         │     Param # │
├──────────────────────────────┼──────────────────────┼─────────────┤
│ lstm_8 (LSTM)                │ (None, 123)          │      61,500 │
├──────────────────────────────┼──────────────────────┼─────────────┤
│ dense_24 (Dense)             │ (None, 64)           │       7,936 │
├──────────────────────────────┼──────────────────────┼─────────────┤
│ dropout_16 (Dropout)         │ (None, 64)           │           0 │
├──────────────────────────────┼──────────────────────┼─────────────┤
│ dense_25 (Dense)             │ (None, 32)           │       2,080 │
├──────────────────────────────┼──────────────────────┼─────────────┤
│ dropout_17 (Dropout)         │ (None, 32)           │           0 │
├──────────────────────────────┼──────────────────────┼─────────────┤
│ dense_26 (Dense)             │ (None, 7)            │         231 │
└──────────────────────────────┴──────────────────────┴─────────────┘

 Total params: 71,747 (280.26 KB)
 Trainable params: 71,747 (280.26 KB)
 Non-trainable params: 0 (0.00 B)
```





3.7 Constraints

Constraints in speech emotion recognition (SER) systems can arise from various factors, impacting their effectiveness and accuracy:

1. Variability in Speech: Natural speech exhibits considerable variability in terms of accent, dialect, speaking rate, pitch, and tone. This variability poses a challenge for SER systems to generalize across different speakers and contexts.

2. Limited Data Availability: Building robust SER models requires large and diverse datasets encompassing various emotions, languages, and speaking styles. However, acquiring such datasets may be challenging due to privacy concerns, data annotation costs, and the need for balanced emotional representations.

3. Ambiguity in Emotional Expression: Emotions are complex and multidimensional, often expressed through subtle variations in speech prosody, intonation, and linguistic content.

Disentangling these subtle cues and accurately recognizing emotions poses a significant challenge for SER systems.

4. Noise and Environmental Factors: Environmental noise, microphone quality, and recording conditions can degrade the quality of speech signals, affecting the performance of SER systems. Robustness to noise and environmental variability is essential for real-world deployment.

5. Cultural and Individual Differences: Emotion expression can vary across cultures and individuals, making it challenging to develop universal SER models that generalize across diverse populations. Cultural norms, social context, and individual personality traits influence how emotions are expressed and perceived.

6. Real-time Processing: In applications requiring real-time emotion recognition, such as human-computer interaction or virtual assistants, SER systems must operate with low latency and high efficiency. Real-time processing constraints impose limitations on model complexity and computational resources.

Addressing these constraints requires interdisciplinary research efforts spanning signal processing, machine learning, psychology, linguistics, and ethics to advance the capabilities and reliability of SER systems for practical applications.

4.conclusion

In this work, a speech emotion recognition system with the LSTM model and MFCC feature is presented. It is observed that MFCC is a popularly used feature and gives better results for emotion detection in SER. There is a future scope in ROC curve improvement. The area observed under the ROC curve is 0.55. 67.21% loss is observed in this model, which needs to improve. The positive point observed in the implemented model is that the system achieves 84.81% accuracy. Still, there is a scope for improvement using a combination of different feature and optimizing ML model for a better true positive rate.

5.References

[1] P. Harár, R. Burget and M. K. Dutta, "Speech emotion recognition with deep learning," 2017 4th International Conference on Signal Processing and Integrated Networks (SPIN), Noida, 2017, pp. 137- 140.

[2] Moataz El Ayadi, Mohamed S. Kamel, Fakhri Karray, Survey on speech emotion recognition: Features, classification schemes, and databases, Pattern Recognition, Vol. 44, Issue 3, 2011.Pages 572-587.

[3] Supriya B.Jagtap, Dr.K.R.Desai, Ms. J. K. Patil ” A Survey on Speech Emotion Recognition Using MFCC and Different classifier”, 8th national conference on emerging trends in engg and technology, 10th march 2018.

[4] Dhruvi desai “Emotion recognition using Speech Signal: A Review”, International Research Journal of Engineering and Technology (IRJET),Volume: 05 Issue: 04 , Apr-2018