**A**

**MAJOR-PROJECT-REPORT**

**ON**

“SPEECH EMOTION RECOGNITION”

Submitted in partial fulfillment of the requirements for the 8th Semester

**Bachelor of Engineering**

**in**

**INFORMATION SCIENCE AND ENGINEERING**

**of**

**Visvesvaraya Technological University, Belagavi**

***By***

**Ms. Arpitha Pradeep (1SI18IS009)**

**Ms. Swathi H. L (1SI18IS052)**

**Mr. Ujwal Patil (1SI19IS405)**

**Mr. Yashvanth. D (1SI18IS062)**

Under the guidance of

Dr. Jayanna H S

Professor,

Dept.of ISE, SIT,

Tumakuru



**Department of Information Science and Engineering,**

**Siddaganga Institute of Technology, Tumakuru-572103**

**2021-22**

SIDDAGANGA INSTITUTE OF TECHNOLOGY

TUMAKURU-572103

**Department of Information Science and Engineering**



**CERTIFICATE**

*This is to certify that the major project entitled* “*Speech emotion recognition*” *has been successfully carried out by Arpitha pradeep bearing USN 1SI18IS009, Swathi. H. L bearing USN 1SI18IS052, Ujwal Patil bearing USN 1SI19IS405 and Yashvanth D bearing USN 1SI18IS062 at* Siddaganga Institute of Technology, Tumakuru, *in partial fulfillment of the requirements for 8th Semester of Bachelor of Engineering in Information Science and Engineering of Visvesvaraya Technological University, Belagavi, during the academic year 2021-22. The project has been approved as it satisfies the academic requirements in respect of Project work prescribed for the Bachelor of Engineering Degree.*

Signature of Guide Signature of HOD

**Dr. H. S Jayanna, Dr. R. Aparna,**

**Professor, Professor and Head,**

**Dept. of ISE, Dept. of ISE,**

**SIT, Tumakuru SIT, Tumakuru**

**External Viva**

**Name of the Examiners Signature with Date**

**1. \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**2. \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**ACKNOWLEDGEMENT**

With reverential pranams, we express our sincere gratitude and salutations to His Holiness **Dr. Sree Sree Sivakumara Swamigalu,** Founder President ofSree Siddaganga Education Society and **Sree Sree Siddalinga Swamigalu**, President of Sree Siddaganga Education Society for their unlimited blessings.

First and foremost, we wish to express our deep sincere feelings of gratitude to our institution, **Siddaganga Institute of Technology,** for providing us an opportunity for completing our project successfully.

We are grateful to **Dr. M N Channabasappa**, Director, and **Dr. Shivakumaraiah**, CEO, Siddaganga Institute of Technology, Tumakuru for their kind cooperation and encouragement.

We express our kind thanks to **Dr. Dinesh S V**, Principal, Siddaganga Institute of Technology, Tumakuru for his encouragement and support to complete our major project.

We express our heartfelt thanks to **Dr. R Aparna,** Professor and Head, Department of Information Science and Engineering, Siddaganga Institute of Technology, Tumakuru for her suggestions and advice.

We express our gratitude and humble thanks to our project guide, **Dr H S Jayanna**, Professor, Department of Information Science and Engineering, Siddaganga Institute of Technology, Tumakuru for guiding and facilitating to complete our Major-project work successfully.

We are conscious of the fact that we have received cooperation in many ways from the **Teaching, Technical and supporting staffs** of the Department of Information Science and Engineering and we are grateful to all for their cooperation.

We express heartfelt gratitude to our **parents and friends** for their constant moral support and encouragement throughout this work.

**- Arpitha Pradeep**

**- Swathi H L**

**- Ujwal Patil**

**-Yashvanth D**

**TABLE OF CONTENT**

[Chapter 1 1](#_Toc107838625)

[INTRODUCTION 1](#_Toc107838626)

[1.1 Objectives 1](#_Toc107838627)

[1.2 Methodology 2](#_Toc107838628)

[Chapter 2 3](#_Toc107838629)

[LITERATURE SURVEY 3](#_Toc107838630)

[Chapter 3 6](#_Toc107838631)

[REQUIREMENT SPECIFICATION 6](#_Toc107838632)

[Chapter 4 7](#_Toc107838633)

[SYSTEM DESIGN 7](#_Toc107838634)

[Chapter 5 11](#_Toc107838635)

[SYSTEM IMPLEMENTATION 11](#_Toc107838636)

[5.1 Data Collection 11](#_Toc107838637)

[5.1.1 RAVDEES (Ryerson Audio-Visual Database of Emotional Speech and Song) 11](#_Toc107838638)

[5.1.2 TESS (Toronto Emotional Speech Set) 11](#_Toc107838639)

[5.1.3 CREMA (Crowd-Sourced Emotional Multimodal Actors) 12](#_Toc107838640)

[5.1.4 SAVEE (Surrey Audio-Visual Expressed Emotion) 12](#_Toc107838641)

[5.2 Data preparation 12](#_Toc107838642)

[5.3 Data Pre-processing 13](#_Toc107838643)

[5.3.3 Trim 14](#_Toc107838644)

[5.3.4 Padding 14](#_Toc107838645)

[5.3.5 Noise 15](#_Toc107838646)

[5.4 Feature Extraction 15](#_Toc107838647)

[5.4.1 Mel-frequency cepstral coefficients (MFCC): 15](#_Toc107838648)

[5.4.2 Zero-Crossing Rate (ZCR): 15](#_Toc107838649)

[5.4.3 Root mean square (RMS): 16](#_Toc107838650)

[5.5 Recurrent Neural Network (RNN) 16](#_Toc107838651)

[5.5.1 Working of RNN 17](#_Toc107838652)

[5.5.2 Training A Recurrent Neural Network 19](#_Toc107838653)

[Chapter 6 21](#_Toc107838654)

[RESULT AND ANALYSIS 21](#_Toc107838655)

[Chapter 7 24](#_Toc107838656)

[CONCLUSION 24](#_Toc107838657)

[BIBLIOGRAPHY 25](#_Toc107838658)

**LIST OF FIGURES**

[Figure 1. Block diagram of SER 7](file:///C:\Users\Admin\OneDrive\Desktop\SPDY\SIT\MAJOR\Reports\Paper\Report%20final.docx#_Toc107838659)

[Figure 2. Flow diagram for SER 8](file:///C:\Users\Admin\OneDrive\Desktop\SPDY\SIT\MAJOR\Reports\Paper\Report%20final.docx#_Toc107838660)

[Figure 3. Flow of Data pre-processing 9](#_Toc107838661)

[Figure 4. Process of feature Extraction 9](#_Toc107838662)

[Figure 5. Flowchart for SER 10](file:///C:\Users\Admin\OneDrive\Desktop\SPDY\SIT\MAJOR\Reports\Paper\Report%20final.docx#_Toc107838663)

[Figure 6. Raw input voice data 13](file:///C:\Users\Admin\OneDrive\Desktop\SPDY\SIT\MAJOR\Reports\Paper\Report%20final.docx#_Toc107838664)

[Figure 7. Optimized voice data 13](file:///C:\Users\Admin\OneDrive\Desktop\SPDY\SIT\MAJOR\Reports\Paper\Report%20final.docx#_Toc107838665)

[Figure 8. Trimmed voice data 14](file:///C:\Users\Admin\OneDrive\Desktop\SPDY\SIT\MAJOR\Reports\Paper\Report%20final.docx#_Toc107838666)

[Figure 9. Padded voice data 14](file:///C:\Users\Admin\OneDrive\Desktop\SPDY\SIT\MAJOR\Reports\Paper\Report%20final.docx#_Toc107838667)

[Figure 10. Noise reduced voice data 15](file:///C:\Users\Admin\OneDrive\Desktop\SPDY\SIT\MAJOR\Reports\Paper\Report%20final.docx#_Toc107838668)

[Figure 11. LSTM Gates Representation 18](#_Toc107838669)

[Figure 12. States of LSTM 19](#_Toc107838670)

[Figure 13. Loss function and Model accuracy graph for 500 epochs 21](file:///C:\Users\Admin\OneDrive\Desktop\SPDY\SIT\MAJOR\Reports\Paper\Report%20final.docx#_Toc107838671)

[Figure 14. Loss function and Model accuracy for 100epochs 21](file:///C:\Users\Admin\OneDrive\Desktop\SPDY\SIT\MAJOR\Reports\Paper\Report%20final.docx#_Toc107838672)

[Figure 15. Confusion matrix for tested data 22](file:///C:\Users\Admin\OneDrive\Desktop\SPDY\SIT\MAJOR\Reports\Paper\Report%20final.docx#_Toc107838673)

[Figure 16. summary report for real time prediction using bar graph. 23](file:///C:\Users\Admin\OneDrive\Desktop\SPDY\SIT\MAJOR\Reports\Paper\Report%20final.docx#_Toc107838674)

**ABSTRACT**

Emotion plays a crucial role in day-to-day interpersonal human interactions. Recent findings have suggested that emotion is integral to rational and intelligent decisions. It helps us to relate with each other by expressing our feelings and providing feedback. This important aspect of human interaction needs to be considered in designing human-machine interfaces (HMIs). To build interfaces that are more in tune with the user’s needs and preferences, it is essential to study how emotion modulates and enhances the verbal and nonverbal channels in human communication. In this project work, we developed a Speech Emotion Recognition(SER) system using MFCC(Mel frequency Cepstral Coefficient), ZCR(Zero Crossing Rate), and RMS(Root Mean Square) for extracting the features and for classifying the emotions into their respective classes, we used RNN(Recurrent Neural Network) classifier.

# Chapter 1

## INTRODUCTION

Emotion is such a unique power of human trial that plays a vital role in distinguishing human civilization from others. Emotion helps us to understand each other in a better way. To express emotions, people use body postures, facial expressions, and vocalizations. Speech is one of the most important media for expressing emotion. We can identify many types of emotions by talking or listening to others. Happiness, sadness, fear, disgust, surprise, and anger are the six universal emotions. Though people use a variety of means to express emotions, the easiest and most complete way to express emotion and feelings is through speech.

In terms of science, Speech is a complex signal consisting of various information, such as messages to be communicated, speaker, language, region, emotions etc. In speech emotion recognition, the emotional state of a speaker is extracted from his or her speech. The acoustic characteristic of the speech signal is Feature and emotions are mainly distinguished based on their energy levels, so along with the acoustic features energy parameters are also considered. One of the most common and efficient methods for feature extraction is the Mel frequency Cepstral Coefficient (MFCC). The objective of Speech emotion Recognition is to pre-process the input signal, extract the required features, and characterize and recognize the speaker's emotions.

We define an Emotion recognition system as a collection of methodologies that process and classify speech 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, conducting therapies, music players, and more.

### Objectives

Speech emotion recognition (SER) can be implemented for human-like behavior in robots which may help to improve the services and, in the field of medical science SER plays an important role. During the pandemic, we have seen the very huge impact of technology which made it possible to virtualize our routine. Likewise, now we have applications to get doctor’s appointments, virtual meetings with doctors, and many others. SER can be helpful in these applications to understand the patient's emotions and for keeping their good mental health while conducting mental therapies to examine the patient's actual mindset and emotions. Thus, as the world is moving towards the growing technology and virtualization with the improvisation and inclusion of robots in day-to-day life, SER plays a major role where human-machine interactions are involved.

In this project to accomplish the need of a system like Speech Emotion Recognition, first we need to collect the standard and required speech data which will be helpful in training the model. And the collected data must be preprocessed to make it easier to interpret and use. The preprocessed data will be used to extract the underlying features by using righteous methods. As we know once the features are extracted, the data is passed into the trained model to analyze and classify the speech data.

### Methodology

In this project, Data Preprocessing is done through pyDub library. Using Librosa library the required features are extracted and for the classification algorithm we have used RNN model along with its LSTM property. The project is built and deployed using Django framework.

Graphical user interface, diagram

Description automatically generated

Figure 1. Block diagram of the project

In the Figure 1 describes the block diagram of the project. Initially, the input speech will be taken in the form the user taken a recording voice data or a single input mp3 file. This speech is fragmented into the number of ‘n’ fragments based on the Window size and sample rate. These fragments are optimized and pre-processed to extract the features. Each fragment is treated as a small input audio file. These extracted features are passed to a trained and tested RNN model which will output the frequency of each emotion. Based on the highest frequency, the emotion of the speech is obtained.

# Chapter 2

## LITERATURE SURVEY

Speech emotion recognition (SER) is one of the challenging topics for Human-Computer Interaction (HCI). In the past years, research has focused on finding the best method for feature extraction and the best model to design and train to get an efficient result. There are some of the existing systems that have provided good insights about SER.

**Badsha and Islam** [1], In this experiment, the authors processed the voice signal through wavelet analysis. They considered three emotions such as joy, sorrow and anger. They used four different persons' speeches in each of the mentioned emotions. A Wavelet is an oscillation that decays quickly The Haar Wavelet is a function that satisfies the conditions of a wavelet. The Haar wavelet was used to process the voice signal since it is well time localized. To predict the approximate emotion the authors used four basic features of voice signal which are mean frequency, max frequency, L1 form and L2 form. After calculating the mentioned features for each audio signal, it has been observed that mean frequency and maximum frequency gradually increase from sorrow, and joy to anger. L1 norm and L2 norm gradually decrease respectively from sorrow, joy to anger.

**Koduru Valveti, and Budati** [2]**,** compared three machine learning Models, Support Vector Machine (SVM), Decision Tree and Linear Discriminant Analysis (LDA) to determine the accuracy of the emotional state of the speech signal. In this paper the standard Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) database is used to recognize the emotions. Butterworth and Chebyshev ﬁlters are used to remove the noise from the speech samples and the cepstral features such as the MFCCs, Pitch, Energy, zero crossing rate (ZCR), Discrete wavelet transform are used for feature extraction. From this paper the Decision tree gives more accuracy approximately to 85% and the SVM yields 70%, and the LDA 65%.

**Ingale and Chaudhari** [3], compare different classification schemes, such as K-Nearest Neighbors (KNN), Hidden Markov Model (HMM), Support Vector Machine (SVM), Artificial Neural Network (ANN) and Gaussian mixed model (GMM). For feature extraction, such as energy, pitch, linear predictive cepstral coefficient (LPCC) and MFCC algorithm are used, and a forward selection algorithm is used for feature selection. The authors conclude that GMM is only best suited to extracting the overall features of the speech signal with an accuracy of 78.77%. ANN can find nonlinear boundaries separating emotional states and has an accuracy of 51.19%. KNN gave an emotion recognition rate of 64% and HMM only matched spectral features with an accuracy of 76.12%. SVM gives a better recognition rate than other classification schemes with an accuracy of 77%. The authors do not emphasize signal based on noise, and all real-time audio samples cannot resist noise.

**Selvaraj** [4]: Shows superior features such as fundamental frequency, loudness, pitch, speech intensity, and glottic parameters, and uses MFCC for spectral feature extraction. The Radial Basis Function (RBF) and Backpropagation Network (BPN) of the classification scheme are used to recognize emotions. And the standard database Berlin (EMO-DB) is used to recognize emotions. Pitch and SVM are used to classify the gender of the speaker. The results show that radial basis functions provide more efficient results than backpropagation networks (82%). The authors were able to find the best function for classifying noise-free signals. They also claimed that the test sample was noise-free.

**Shambhavi and Nitnaware** [5]**,** proposed a feature extraction algorithm that uses MFCCs and SVMs to classify emotions. The working paper also highlights emotion recognition and industrial applications, where this system is needed, such as call centers to improve emotional states depending on the situation, human-robot interfaces, and intelligent voice instruction systems. The author neglected to report the accuracy report of the implemented GUI.

**Arias** [6]**,** proposed a new shape-based method by using a neutral model to detect the emotional enhancement of the fundamental frequency. The new method is supported by Functional Data Analysis (FDA), which aims to maintain natural variation in F0 contours. For a given F0 contour, the PCA is calculated for use as a feature of voice emotion detection. Empirical results show that the approximation accuracy proposed in the binary classification reaches 75.8%. This is 6.2% more than a trained benchmark system with overall F0 static. This approach is evaluated by the SEMAINE dataset. The results show that for speech emotion recognition, using shape-based methods may be effective in real-world applications.

# Chapter 3

## REQUIREMENT SPECIFICATION

##### 3.1 Python

Python is an easy to adjust, mind boggling programming language. It has capable unusual state data structures and a clear yet fruitful approach to manage thing arranged programming. python's rich etymological structure and dynamic forming, together with its deciphered nature, make it an ideal language for scripting and quick application. There are some modules that were utilized in this venture, every one of them being python libraries, some of them utilized are referenced beneath.

##### 3.2 Django

Django is a high-level Python web framework that enables rapid development of secure and maintainable websites. Built by experienced developers, Django takes care of much of the hassle of web development, so we can focus on writing your app without needing to reinvent the wheel. It is free and open source, has a thriving and active community, great documentation.

##### 3.3 Tensor flow

TensorFlow provides pre-built functions and advanced operations to ease the task of building different neural network models. It provides the required infrastructure and hardware which makes them one of the leading libraries used extensively by researchers and students in the deep learning domain.

##### 3.4 Keras

Keras has strong multi-GPU & distributed training support. Keras is scalable. Using the TensorFlow Distribution Strategy API, which is supported natively by Keras, you easily can run your models on large GPU clusters (up to thousands of devices) or an entire TPU pod, representing over one exaflops of computing power.

# Chapter 4

## SYSTEM DESIGN

Speech Emotion Recognition build with python language using Django framework. Django is a trending framework for construct web-based projects. SER front end static pages are handled by the HTML, CSS, and JS. The back-end control is taken by the python libraries like librosa, pyaudio, PyDub and Django web handler. SER includes several data pre-processing steps and the pre processed data will use to extract the features. MFCC, ZCR, and RMS are the main features to extract the energy level of the audio sample.

### 4.1 Flowchart

Diagram

Description automatically generated

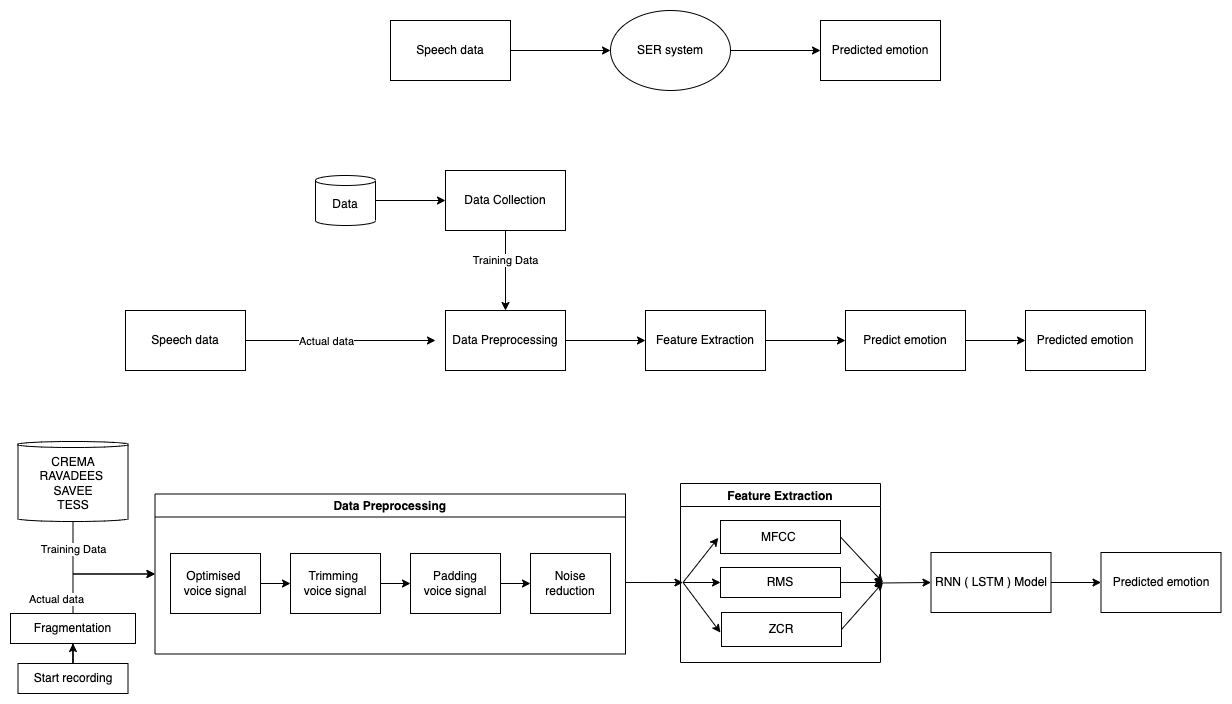
Figure . Flow diagram for speech recognition

Figure 2 shows the flow of data in SER. The input will be captured when the recording of speech starts, and the model starts to read the data. As the reading of speech continues, it will be fragmented into number of fragments every 15 second. These fragments undergo data preprocessing, which further is used for extracting the required features for predicting the correct emotion. This happens with number of different chunks by predicting the emotion in each chunk, later which will integrated and the emotion with highest probability is outputted. When the recording stops, the final output is displayed as the recognized emotion in the input speech.

Diagram

Description automatically generatedIn the { } shows the Flowchart for Speech emotion Recognition. SER is maintained with the two threads, one thread T2 handles the frontend part and another Thread T1 handles the backend part of the project. T1 thread is used to listen to the client's voice and meanwhile it will break down and save the continuous voice input into small voice fragments of 5 seconds in length. Once the voice fragment is saved in the temporary folder it is ready for data pre-processing. All 3-feature extraction (MFCC, RMS, ZCR) is done by taking processed data. The features are given to the RNN model to predict the emotion. Once the user clicks the stop recording thread T1 collects all fragments prediction results, and it gives all frequency of emotions.

### 4.2 Data Flow Diagram

The above {} shows the Level 0 Data flow diagram, also known as Context diagram of SER system. It represents the entire SER system and input/output data. Input data is representing the user audio file and the output represents the emotion of the given input audio file.

{ } represents the more detailed processing, this is called as Level 1 Data Flow Diagram. As we can observe the processes such as data collection, data preprocessing, Feature extraction and Emotion prediction. The collected data is trained and preprocessed. The speech data taken from user is also preprocessed and subjected to extract features which will be passed into trained model for emotion prediction.

Graphical user interface, diagram, application, Teams

Description automatically generated

{ } shows the level 2 Data flow diagram with much detailed sub-processes. To train the data, we have selected a vast databases such as RAVDESS, TESS, CREMA and SAVEE. The user will start recording the input speech, it will be fragmented every 15 second. This data is subjected to data preprocessing. Initially we will take raw data which is the voice signal taken from the user directly. then it will be converted into normalised data which uses the signal sequence divided by the highest value of the signal to ensure that each sentence has a comparable volume level. After data normalization, the data will be trimmed because it will control signals and removes the spaces between word to word. Once the trimmed data is obtained padding is done, this will reduce the signal levels before the active amplification process to avoid overloading the microphone circuitry. Later the noise from the signal is removed.

In the next step the preprocessed data is used to extract the features underlying. At the brgining, the data is divided into number of chunks where a chunk is a second level fragmentation product. From these chunks the required features will be extracted through series of process like MFCC(Mel frequency Cepstral coefficient), ZCR(Zero Crossing Rate) and RMS(Root Mean Square). These extracted features are passed to the trained model.The model used in the project is RNN(Recurrent Neural Network) with LSTM(Long-Short Term Memory) features. The model will analyze the pattern in the subjected data and predict the emotion in it. This predicted emotion is shown to user as the result.

### Graphical user interface, diagram Description automatically generated4.3 Use case diagrams

# Chapter 5

## SYSTEM IMPLEMENTATION

### 5.1 Data Collection

Training a model is also one of the important steps in the process. To train a model requires suitable data. In this project TESS (Toronto Emotional Speech Set), CREMA (Crowd-Sourced Emotional Multimodal Actors), SAVEE (Surrey Audio-Visual Expressed Emotion), and RAVDESS (Ryerson Audio-Visual Database of Emotional Speech and Song) are considered. Each dataset contains a minimum of 6 different emotions. Each dataset contains audio files which express each emotion by considering various scenes. There are almost 8 different emotions, ~500 Different Scenes, 34 different Actors, and a total of 14962 files are used to create the customized dataset.

#### 5.1.1 RAVDEES (Ryerson Audio-Visual Database of Emotional Speech and Song)

RAVDEES is a validated multimodal database of emotional language and songs. The database consists of 24 professional actors vocalizing lexically matched utterances in a neutral North American accent with gender balance. Voices include expressions of calm, happiness, sadness, anger, fear, surprise, and disgust, and songs contain emotions of calm, happy, sad, anger, and fear. Each expression produces two levels of emotional intensity and neutral expression. All conditions are available in face and voice, face-only, and voice-only formats. The set of 7356 recordings, each rated 10 times on emotional validity, intensity, and verity. Ratings were provided by 247 untrained research participants from North America. An additional set of 72 participants provided test-retest data. High levels of emotional validity and test-retest concordance reliability were reported.

#### 5.1.2 TESS (Toronto Emotional Speech Set)

This dataset would be in the service of a very good training dataset for the emotion classifier in terms of generalization. The dataset consists of 200 target words spoken by two actresses and recordings were made of the set portraying each of seven emotions (disgust, happiness, surprise, sadness, anger, fear, and neutral). There is a total of 2800 audio files. The dataset is organized such that each of the two female actors and their emotions is contained within its own folder. And within that, all 200 target word audio files can be found. The audio file is in the form of a WAV format.

#### 5.1.3 CREMA (Crowd-Sourced Emotional Multimodal Actors)

CREMA dataset is a very good dataset to use to ensure the model does not overfit. This dataset 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 origins and ethnicities. Actors spoke from a selection of 12 sentences. The sentences were presented using one of six different emotions and four different emotion levels.

#### 5.1.4 SAVEE (Surrey Audio-Visual Expressed Emotion)

This dataset was recorded from four native English male speakers, postgraduate students, and researchers at the University of Surrey aged from 27 to 31 years. Emotion has been described psychologically in discrete categories: anger, fear, disgust, surprise, happiness, and sadness. We attached neutral to provide recordings of 7 emotion categories. The text material consisted of 15 sentences per emotion: 3 common, 2 emotion-specific, and 10 generic sentences that were different for each emotion and phonetically balanced. The 3 common and 2\*6 = 12 emotion-specific sentences were recorded as neutral to give 30 neutral sentences.

### 5.2 Data preparation

A continuous speech signal may contain surrounding disturbance, silence or sometimes the same speech segments overlapped with others. This may increase the complexity of speech processing. So, the speech signal is split into short-time frames because the frequency of the signal changes over time and it will be difficult for transforming from the time domain to the frequency domain, to get individual frequency signals [5]. Thus, the taken frame length is 2048 samples. Hop length is the length of the non-intersecting portion in the audio fragment [6]. Hop length considered here is 512 samples. These two parameters are necessary for obtaining detailed processing of audio signals.

Total audios hops =

=

### 5.3 Data Pre-processing

Before the extraction of the features of the user voice input, this voice data is manipulated by using pre-processing methods.

#### **5.3.1 Raw data**

Raw data is the voice signal taken from the user directly. This voice data includes many unwanted and lack of usable precise data. Sample of the raw data signal shown in the Figure 6.

A picture containing text

Description automatically generated

Figure . Raw input voice data

#### **5.3.2 Normalized signal**:

It is a strategy for modifying the volume of sound to a standard level. Normalization is used the signal sequence divided by highest value of the signal to ensure that each sentence has a comparable volume level. Optimized voice data is shown in the Figure 7.

**A picture containing text

Description automatically generated**

Figure . Optimized voice data

Audio files where the speech is loud in some portions and quiet in others. Having this variance in volume can hinder transcription. Luckily, PyDub's effects module has a function called normalize() which finds the maximum volume of an AudioSegment, then adjusts the rest of the AudioSegment to be in proportion. This means the quiet parts will get a volume boost. headroom is how close to the maximum volume to boost the signal up to (specified in dB).

#### **5.3.3 Trim**

Trim refers to changing the level of a signal (up or down) to align it more appropriately for a particular device. A trim control may apply some “gain” to raise a signal level, or it may apply some negative gain, or attenuation to lower a signal. Trimmed voice data is shown in the Figure 8.

Figure . Trimmed voice data

A picture containing text, monitor, aircraft, silhouette

Description automatically generated

#### **5.3.4 Padding**

Padding reduce signal levels before the active amplification process to avoid overloading the microphone circuitry. Figure 9 shows the padded data to the trimmed voice data.

Shape, rectangle

Description automatically generated

Figure . Padded voice data

#### **5.3.5 Noise**

Noise reduction is the process of removing noise from a signal. Noise reduction techniques exist for audio and images. Noise reduction algorithms may distort the signal to some degree. All signal processing devices, both analog and digital, have traits that make them susceptible to noise. Figure 10 shows the noise reduced data.

Shape, rectangle

Description automatically generated

Figure . Noise reduced voice data

### 5.4 Feature Extraction

#### **5.4.1 Mel-frequency cepstral coefficients (MFCC):**

MFCC is a speech emotion feature parameter that is an inverse spectral coefficient extracted in the frequency domain of the Mel scale, a feature widely used in automatic speech and speaker recognition. The Mel scale is very accurate in describing the nonlinear characteristics of human ear frequency. The computational relationship between it and the frequency can be expressed in 5.4.1.

… (5.4.1)

Where :  
🡪 Meier frequency function.  
 🡪 linear frequency

#### **5.4.2 Zero-Crossing Rate (ZCR):**

As the name suggests the zero-crossing rate is the measure of the rate at which the signal is going through the zeroth line more formally signal is changing from positive to negative or vice versa. Equation 5.4.2 Mathematically it can be measured as.

… (5.4.2)

Where   
 🡪 signal  
🡪 length of signal

#### **5.4.3 Root mean square (RMS):**

The square root of the mean of the square. RMS is a meaningful way of calculating the average of values over a period. With audio, the signal value (amplitude) is squared and averaged over a period, then the square root of the result is calculated. The result is a value, that when squared, is related (proportional) to the effective power of the signal. RMS of a signal is a just a value used to calculate average, or continuous, power.

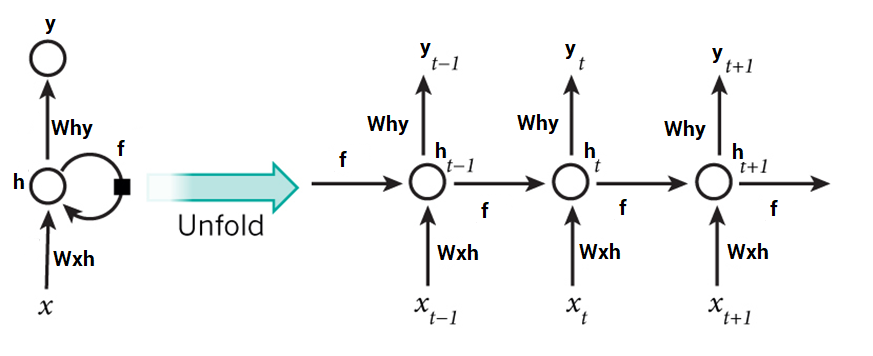
Unfortunately, calculating the RMS value of anything but a simple sine wave (.707 of peak) is very difficult. The further a signal gets in harmonic content from a sine wave, the less accurate RMS values will be. It is nearly impossible to get even close to a true RMS value for a dynamic signal like most music.

… (5.4.3)

### 5.5 Recurrent Neural Network (RNN)

RNN - Recurrent Neural Network is one of the most common Neural networks used for Sequential data. As we know the audio signal is continuous or sequential data, Output for a given input is based on the input at a given time and the previous input in hidden layers. RNN as the name refers to is recurrent, it has a recurrent connection to the hidden state. This looping constraint ensures that sequential information is captured in input data. LSTM (Long Short-Term Memory) keeps the information of current input and previous input for future calculations which helps in getting a pattern from sequential data.

#### 5.5.1 Working of RNN



Initially input is provided to the network which is a single-time step in the process and let's consider the input as ‘’ at the time ‘’. Now the current state () calculation takes place by the combination of the current input and the previous state.

 ...(Eq. 5.5.1.1)

where,

🡪 Weight of the previous state,

  🡪 previous state,

🡪 weight of the current input.

From the above eq.(5.5.1.1), the recurrence relation is shown by the consideration of the previous state to calculate the current state. The Recurrent neuron takes the immediate previous state into consideration. For longer sequences, the equation can involve multiple such states. Once the final state is calculated the output( ) is produced.

… (Eq. 5.5.1.2)

where,  
 🡪 weight of the final state,  
 🡪final state.

The obtained output (yt) is then compared with the actual output, this step will provide the difference between actual output and obtained output which is known as error. The error generated is propagated back to the network to update the weights.

The Activation Function used here is softmax activation function in the recurrent neural network.

##### LSTM

Long Short-Term Memory Network is an advanced RNN, a sequential network, that allows information to persist. It can handle the vanishing gradient problem faced by RNN. A recurrent neural network is also known as RNN is used for persistent memory.

###### **LSTM Architecture**

At a high-level LSTM works very much like an RNN cell. Here is the internal functioning of the LSTM network. The LSTM consists of three parts, as shown in the image below and each part performs an individual function. The first part chooses whether the information coming from the previous timestamp is to be remembered or is irrelevant and can be forgotten. In the second part, the cell tries to learn new information from the input to this cell. At last, in the third part, the cell passes the updated information from the current timestamp to the next timestamp.

These three parts of an LSTM cell are known as gates. The first part is called Forget gate, the second part is known as the Input gate and the last one is the output gate.

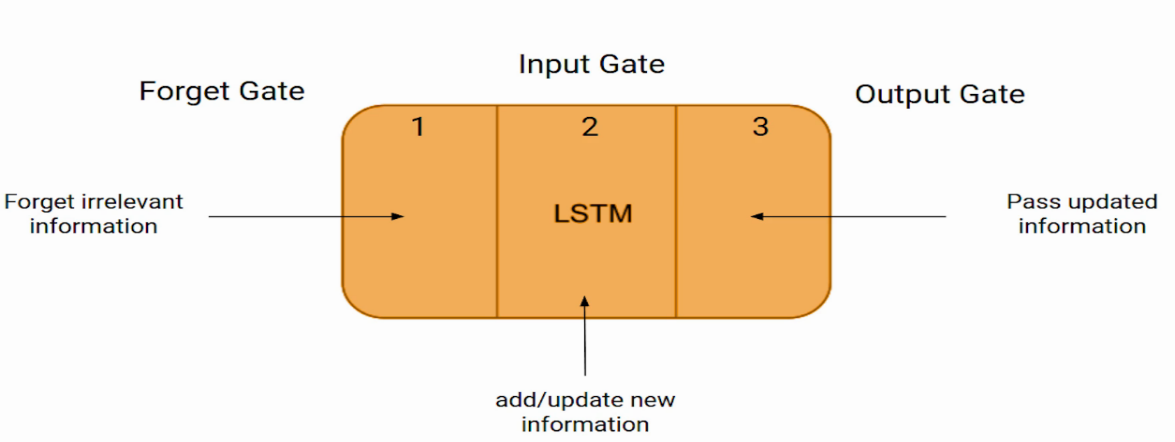


Figure . LSTM Gates Representation

Just like a simple RNN, an LSTM also has a hidden state where represents the hidden state of the previous timestamp and Ht is the hidden state of the current timestamp. In addition to that LSTM also have a cell state represented by and for previous and current timestamp respectively.

Here the hidden state is known as short term memory and the cell state is known as long term memory. Refer to the following image Figure 12.

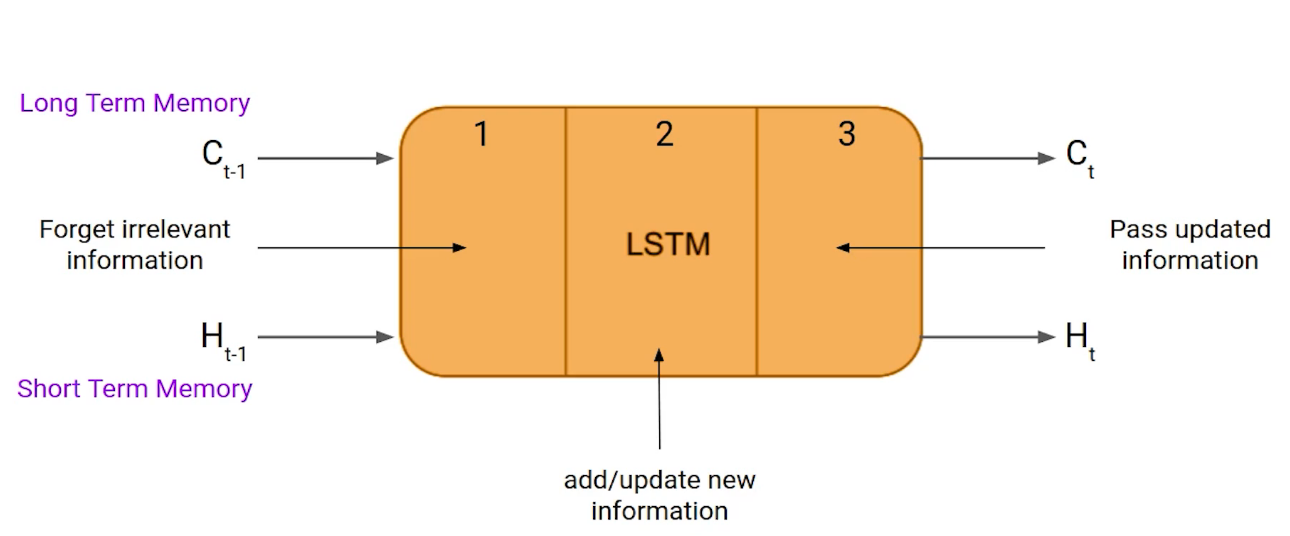


Figure . States of LSTM

It is interesting to note that the cell state carries the information along with all the timestamps.

#### 5.5.2 Training A Recurrent Neural Network

The backpropagation algorithm of an artificial neural network is modified to include the unfolding in time to train the weights of the network. This algorithm is based on computing the gradient vector and is called back propagation in time or BPTT algorithm for short. The pseudo-code for training is given below. The value of can be selected by the user for training. In the pseudo-code below pt is the target value at time step :

Repeat till stopping criterion is met:

Set all to zero.

Repeat for to

Forward propagate the network over the unfolded network for time steps to compute all and .

Compute the error as:

Backpropagate the error across the unfolded network and update the weights.

LSTM were also designed to address the vanishing gradient problem in RNNs. LSTM use three gates called input, output and forget gate. Like GRU, these gates determine which information to retain.

# Chapter 6

## RESULT AND ANALYSIS

Chart, line chart

Description automatically generatedChart

Description automatically generatedAfter configuring the model, given a shot for the higher value of epoch, which gives an overall picture of the model towards overfitting and underfitting. The following graph Figure 13 of training accuracy vs. validation accuracy over the 500 epochs.

Figure . Loss function and Model accuracy graph for 500 epochs

Chart, histogram

Description automatically generatedA picture containing chart

Description automatically generatedAs seen in the graph the model will get overfitted after 100 epochs. To avoid overfitting, the model must be trained with the epoch value less than or equal to 100. The following graph Figure 14 of training accuracy vs. validation accuracy over the 100 epochs.

Figure . Loss function and Model accuracy for 100epochs

Table

Description automatically generated

Figure . Confusion matrix for tested data

These results Figure 15 show that our model can distinguish angry emotions better than other emotions. In the above confusion matrix, we can see 9 Surprised emotions get predicted as Happy, which in a way indicates that the model takes the loudness of speech to detect emotion into account. When it comes to Neutral and sad emotions, the model gets confused between Neutral, Sad, and calm because they are almost similar in terms of energy and pitch.

A picture containing chart

Description automatically generatedIcon

Description automatically generated

Figure . summary report for real time prediction using bar graph.

Each graph from Figure 16 is the graph of real time investigation through Jupiter notebook, and it represents the emotions present in the fragment of 15 second of input speech and the last graph represents the most predicted emotion in the provided input speech signal. In real time analysis, continuous speech is taken thus it is divided into number of fragments which further divided into number of chunks thus these graphs in Figure 16 represent each voice fragments. In each chunk the probability of emotions is predicted. Likewise, a greater number of predictions is done for overall input speech.

Based on the integration of number of predictions from all the chunks, the emotion of the input speech is predicted precisely as shown in the last graph of Figure 16. From the observation, the model can predict the emotion in the speech precisely when the number of predictions is more.

|  |  |  |
| --- | --- | --- |
| **Method** | **Dataset** | **Model accuracy / Validation accuracy** |
| Simple RNN | CREMA, RAVADEES, SAVEE, TESS | 60/26 |
| CNN | CREMA, RAVADEES, SAVEE, TESS | 58/38 |
| LSTM | CREMA | 65/59 |
| LSTM | RAVDESS | 92/20 |
| LSTM | TESS | 98/36 |
| DNN (LSTM) | RAVDESS, TESS | 60/49 |
| PCA | SAVEE | 30/26 |
| LSTM | CREMA, RAVADEES, SAVEE, TESS | 68/56 |

# Chapter 7

## CONCLUSION

In recent years, SER technology as one of the key technologies in human-computer interaction (HCI) systems, has received a lot of attention from researchers at home and abroad for its ability to accurately recognize emotions and thus improve the quality of human-computer interaction. In this paper, we propose a Recurrent Neural Network (RNN) algorithm with fused features for SER. in terms of data processing, we quadruple processed all four standard datasets RAVDESS, CREMA, TESS, and SAVEE with 14,962 audio samples. For the network structure, we constructed LSTM Recurrent neural networks (LSTM RNN) to extract spatial features and a transform encoder network to extract temporal features to classify emotions from one of the eight categories.

Taking advantage of the simple structure of LSTM RNN in spatial feature representation and sequence coding transformation, we obtained an accuracy of 68% on the holdout test set. Even though the accuracy is under 70% the model can predict the emotion precisely because of the greater number of predictions obtained through all the chunks of input speech.

# Bibliography

|  |  |
| --- | --- |
| [1] | I. Badsha, "Emotion Detection by Analyzing Voice Signal Using Wavelet," vol. 10, no. 4, December 2020. |
| [2] | K. Anusha, V. H. Bindu, Budati and A. Kumar, "Feature extraction algorithms to improve the speech emotion recognition rate," vol. 3, January 2020. |
| [3] | Ingale and D. S. Chaudhari, "Speech emotion recognition," *International Journal of Soft Computing and Engineering,* vol. 2, pp. 235-238, 2012. |
| [4] | M. Selvaraj, R. Bhuvana and &. P. S. , "Human speech emotion recognition," vol. 8, pp. 311-323, 2016. |
| [5] | T. Singh, "MFCC's Made Easy An easy explanation of an important," 2019. [Online]. Available: https://medium.com/@tanveer9812/mfccs-made-easy-7ef383006040. |
| [6] | Maheshwari and Harsh, "Terms you need to know to start Speech Processing with Deep Learning," Towards Data Science, [Online]. Available: https://towardsdatascience.com/all-you-need-to-know-to-start-speech-processing-with-deep-learning-102c916edf62. |
| [7] | J. P. Arias, C. Busso and N. B. Yoma, "Shape-based modeling of the fundamental frequency contour for emotion detection in speech," *Comput. Speech Lang,* vol. 28, p. 278–294, 2014. |
| [8] | S. S. Shambhavi and V. N. Nitnaware, "Emotion speech rec-ognition using MFCC and SVM," *International Journal of Engi-neering Research & Technology,* 2015. |