A

PRELIMINARY PROJECT REPORT

ON

A CLINICAL APPLICATION FOR SPEECH EMOTION RECOGNITION

SUBMITTED TO THE SAVITRIBAI PHULE PUNE UNIVERSITY, PUNE
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE AWARD OF THE DEGREE OF

BACHELOR OF ENGINEERING
INFORMATION TECHNOLOGY

BY

Pushkar Kane	B190058595
Pratik Mathe	B190058635
Yash Waghumbare	B190058747

Under the guidance of
Prof. Shreesudha Kembhavi

Department Of Information Technology
Pune Institute of Computer Technology
Pune - 411 043.

2022-2023

SCTR's PUNE INSTITUTE OF COMPUTER TECHNOLOGY
DEPARTMENT OF INFORMATION TECHNOLOGY

C E R T I F I C A T E

This is to certify that the preliminary project report entitled
A Clinical Application for Speech Emotion Recognition
submitted by

Pushkar Kane	B190058595
Pratik Mathe	B190058635
Yash Waghumbare	B190058747

is a bonafide work carried out by them under the supervision of **Prof. Shreesudha Kembhavi** and it is approved for the partial fulfillment of the requirement of Savitribai Phule Pune University for the award of the Degree of Bachelor of Engineering (Information Technology).

This project report has not been earlier submitted to any other Institute or University for the award of any degree or diploma.

Prof. Shreesudha Kembhavi
Project Guide

Dr. A. S. Ghotkar
HOD IT

SPPU External Guide

Dr. S. T. Gandhe
Principal

Date:
Place:

Acknowledgement

Purpose of acknowledgments page is to show appreciation to those who contributed in conducting this dissertation work / other tasks and duties related to the report writing. Therefore when writing acknowledgments page you should carefully consider everyone who helped during research process and show appreciation in the order of relevance. In this regard it is suitable to show appreciation in brief manner instead of using strong emotional phrases.

In this part of your work it is normal to use personal pronouns like "I, my, me" while in the rest of the report this articulation is not recommended. Even when acknowledging family members and friends make sure of using the wording of a relatively formal register.

The list of the persons you should acknowledged, includes guide (main and second), head of dept, academic staff in your department, technical staff, reviewers, head of institute, companies, family and friends.

You should acknowledge all sources of funding. It's usually specific naming the person and the type of help you received. For example, an advisor who helped you conceptualize the seminar,someone who helped with the actual building or procedures used to complete the seminar,someone who helped with computer knowledge, someone who provided raw materials for the seminar, etc.

Pushkar Kane	B190058595
Pratik Mathe	B190058635
Yash Waghumbare	B190058747

Sponsorship letter (if any)

Abstract

In this age of artificial intelligence, speech-emotion recognition is crucial. AI can free up a doctor's time to devote to providing patients with individualized care and medical services rather than transactional activities. Analysis and classification of speech signals are used in speech emotion recognition to find the underlying emotions. However, due to the complexity of emotions, which makes them challenging to perceive, the existing SER models still do not reliably understand human emotions. Despite the fact that this profession has been present for a while, new developments have brought it back into the spotlight.

centered on use in medical care, which can be utilized to better the medical treatment by relating the relationship between illness and interaction and the sensations that patients and doctors are now experiencing during a visit.

Using Deep Learning techniques, we will create a web application for clinical use that can identify and categorise emotions from human speech in speech and aid in better patient interpretation to improve medical treatment by tying illness and interaction together.

Keywords: Speech Emotion Recognition, Convolutional Neural Network, Deep Learning, MFCC

Contents

List of Figures

Abbreviations

SER	:	Speech Emotion Recognition
CNN	:	Convolutional Neural Network
MFCC	:	Mel frequency cepstral coeffi- cient
DNN	:	Deep Neural Network
LSTM	:	Long short-term memory

1. Introduction

1.1 Introduction

Currently, artificial intelligence (AI) is empowering a variety of medical applications, including precision medicine, breast cancer imaging diagnostics, and healthcare.

There have been efforts made to help machines understand human emotions because they are an essential component of human interactions and help individuals understand one another better.

AI can relieve doctors of the burden of comprehending their patients' emotional states and move the emphasis from transactional chores to individualised medical treatment and service. However, it necessitates that computers cleverly deduce human speech and comprehend it on a semantic level.

Systems for detecting the embedded emotions in voice signals are referred to as Speech Emotion Recognition (SER) systems. The primary goal of SER systems is to identify certain speaker voice traits under various emotional states.

centred on use in medical care, which can be utilised to better the medical treatment by relating the relationship between illness and interaction and the sensations that patients and doctors are now experiencing during a visit.

SER is the quickest means of communication and information exchange between people and computers, and it has a wide range of practical applications in the field of human-computer interaction.

It results in the expanding study area of Speech Emotion Recognition (SER), where many developments could result in enhancements in a number of fields, including spontaneous translation systems, speech-to-text synthesisers, machine-human interaction, etc.

1.2 Motivation

The goal of speech emotion recognition (SER) is to identify an emotion in speech, regardless of the semantic content. However, since emotions are arbitrary, it might be challenging to record them in everyday speech, even for humans.

However, due to the complexity of emotions, which makes them challenging to perceive, the existing SER models still do not reliably understand human emotions.

Despite the fact that this profession has been present for a while, new developments have brought it back into the spotlight. focused on use in medical care, which can be utilised to better the medical care by relating the interaction between sickness and interaction between doctors and patients at the time of a visit.

1.3 Objectives

The improvement of the interface between humans and machines is SER's main goal. It can also be utilised in lie detectors to track a person's psychophysical condition. Speech emotion recognition has recently found use in forensic science and medicine.

Our actions are influenced by our emotions, such as sadness, happiness, and the flight or freeze reaction. People can know a patient is stressed out by their emotions and may need help.

1.4 Scope

Speech emotion recognition will soon be included into clinical perception in order to assist physicians in monitoring patients' conditions while they are receiving therapy.

The project's scope is as follows:

•Real-time software.

•easy-to-use application

•removes human involvement when assessing the patient's mental state.

2. Literature Survey

2.1 Existing Methodologies

Before the era of deep learning, for SER, researchers mostly use complex hand-crafted features and traditional machine learning methods (such as HMM, SVM, etc.) [1]. In 2014, K. Han et al.[2] proposed the first end-to-end deep learning SER model. In this they use deep neural networks (DNNs) to extract high level features from data and show that they are efficient for speech emotion recognition(SER). In this MFCC is used to extract features and audio file is classified into 5 emotions with 48

In[3], Emotion Recognition on The Basis of Audio Signal Using Naive Bayes Classifier(2016), author has considered different have considered different statistical features of pitch, energy, and ZCR (Zero Crossing Rate) MFCC (Mel frequency cepstral coefficient) from 2000 utterances of the created audio signal database. In that, Pitch feature is extracted by AMDF. Naive bayes is used to classify audio signal. In[4] Automatic speech emotion recognition using recurrent neural networks with local attention author used RNN on IEMOCAP dataset and used an raw and emotional LLD's for feature extraction, had an accuracy of an 58

Fatemeh Noroozi [5] proposed an Vocal based emotion recognition(2017) using random forests and decision tree The average recognition accuracy rate was 66.28

In 2017, Speech Emotion Recognition[6] from Spectrograms with Deep Convolutional Neural Network worked on CNN to have The proposed CNN model consists of three convolutional layers, three fully connected layers.

In [7], authors have explored deep learning model that combines temporal and spatial features They have quadrupled the RAVDESS dataset using AWGN (Additive Gaussian White Noise) for 5760 audio samples. They have built two parallel convolutional neural networks (CNN) to extract spatial features and extract temporal features, classifying emotions from one of 8 classes

MFCC is widely used to analyze any speech signal and had performed well for speech-based emotion recognition systems compared to other features. In 2020 ,Mustaqeem [9] proposed CNN model with deep bidirectional LSTM that used MFCC and produced better result than co-existing model on IEMOCAP dataset. Two experiments were carried out by author to check the effectiveness of the state-of-art model of CNN and DSCNN [10] and achieved an accuracy 79.4

2.2 Research Gap Analysis

The attention mechanism was introduced to improve the performance of the encoder-decoder model for machine translation. The idea behind the attention mechanism was to permit the decoder to utilize the most relevant parts of the input sequence in a flexible manner, by a weighted combination of all the encoded input vectors, with the most relevant vectors being attributed the highest weights. Ever since, attention mechanisms have been applied to various machine learning tasks, such as key-term extraction , image classification, etc.. In the speech emotion-recognition task, the attention mechanism can be used to focus the model on the part that can better express the emotional information, to ignore the irrelevant information and to improve the recognition performance.

3. Requirement Specification and Analysis

3.1 Problem Definition

The Problem is to build a clinical application using Deep learning techniques that can detect patient's emotions embedded in speech and help better interpretation of patient's feelings to improve the medical treatment through the relationship between illness and interaction.

3.2 Scope

Speech emotion recognition will soon be introduced in perceptive of clinical approach to help psychiatrist to check patient's condition while he/she went for treatment.

The scope of the project is as below:

- Real-time application.
- User-friendly application.
- Removes the human interference to check patient's mental condition.

3.3 Objectives

- To Propose a model to recognize emotion from speech using the librosa and sklearn libraries and the RAVDESS dataset.
- To provide better services and better human-machine interactions.

3.4 Proposed Methodology

- Proposed system will classify speech signals to detect emotions embedded in them.
- Preprocessing: Removal of noise from the audio input and conversion to .wav format
- Feature Extraction: We use Mel-scale Frequency Cepstral Coefficients (MFCCs) as input, an audio feature that is widely used in the field of speech recognition.
- Building and tuning a Convolutional Neural Network model to classify into 4 different emotions with increased accuracy.

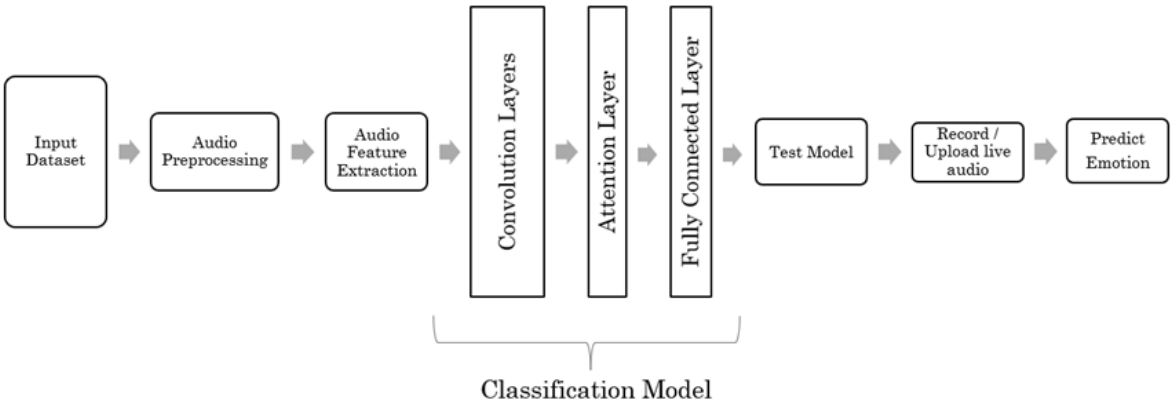


Figure 3.1: The proposed system mode of CNN.

- Compare performances of built model to other coexisting models Distinguish between male and female speech.
- Build a web application for clinical purpose that can take the audio input and predict emotion.

3.5 Project Requirements

3.5.1 Datasets

RAVDESS. Around 1500 audio files from 24 different actors are included in this collection. 12 male and 12 female performers record brief audio clips portraying 8 various emotions, including neutral, calm, happy, sad, furious, afraid, disgusted, and astonished. There are 7,356 files in it. There are two emotional intensity levels (normal and strong) and one neutral expression produced for each expression. Each audio file is titled so that the seventh character corresponds to the many emotions it represents.

3.5.2 Functional Requirements

1. System should throw error in case of wrong input value:If a user gives an incorrect value the system should throw error and make the user to fill the data correctly.
2. System should throw an error if any input is not filled: If a user/patient does not fill all the information required by the model the system should throw an error.

3.5.3 Non Functional Requirements

1. The software must be cross-platform.
2. Each request should be processed within 10 seconds.
3. When there are more than **10000** simultaneous visitors, the website should load in 3 seconds.

3.5.4 Hardware Requirements

Laptop/ Desktop with minimum configuration as:

1. Intel Core i3 1.60 GHz
2. 4 GB RAM
3. 1 TB HDD
4. 15"Color Monitor
5. Keyboard
6. Mouse

3.5.5 Software Requirements

1. Operating System (Any one)
    - Windows
    - Linux
    - Mac OS
2. Python v3 or higher version
3. Jupyter/ Colab Notebook
4. Libraries Used:
    - Pandas
    - Numpy
    - Seaborn
    - Matplotlib
    - Sklearn

3.6 Project Plan

3.6.1 Project Resources

1. **Human Resources**
   Members should have enough knowledge resources: Machine Learning , Deep Learning, and Web Development.

2. **Material Resources**
   Ml IDE- Colab/Jupyter Notebook, Visual Studio

3.6.2 Module Split-up

- Audio Preprocessing
- Feature Extraction
- Model Building
- Predictions on Test data
- Calculating Accuracy (Actual v/s Predicted)
- Live Application

3.6.3 Project Team Role and Responsibilities

- Contributing to overall project objectives
- Completing individual deliverables
- Providing expertise
- Working with users to establish and meet project needs
- Documenting the process

3.6.4 PERT Diagram

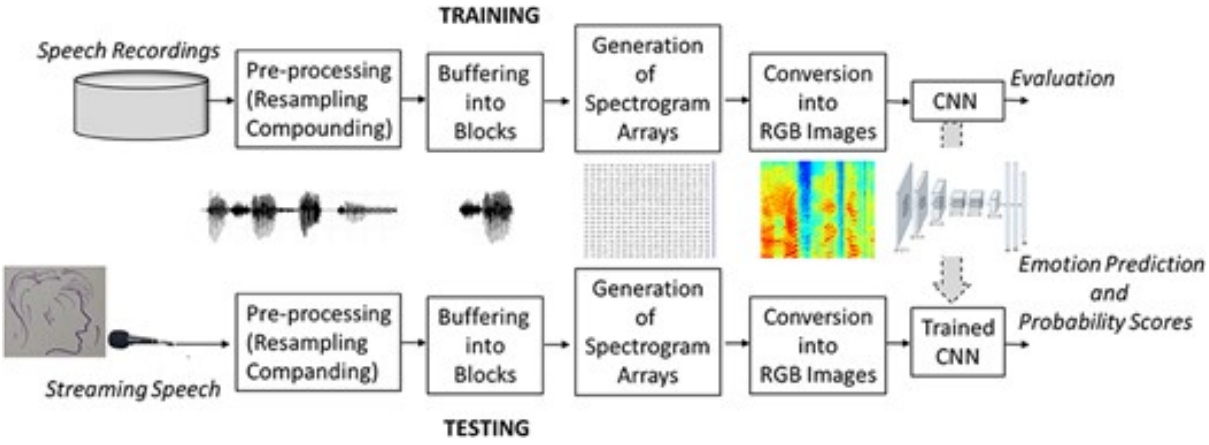


Figure 3.2: PERT diagram

4. System Analysis and Design

4.1 System Architecture

4.2 Necessary UML Diagrams

4.2.1 Use Case Diagram

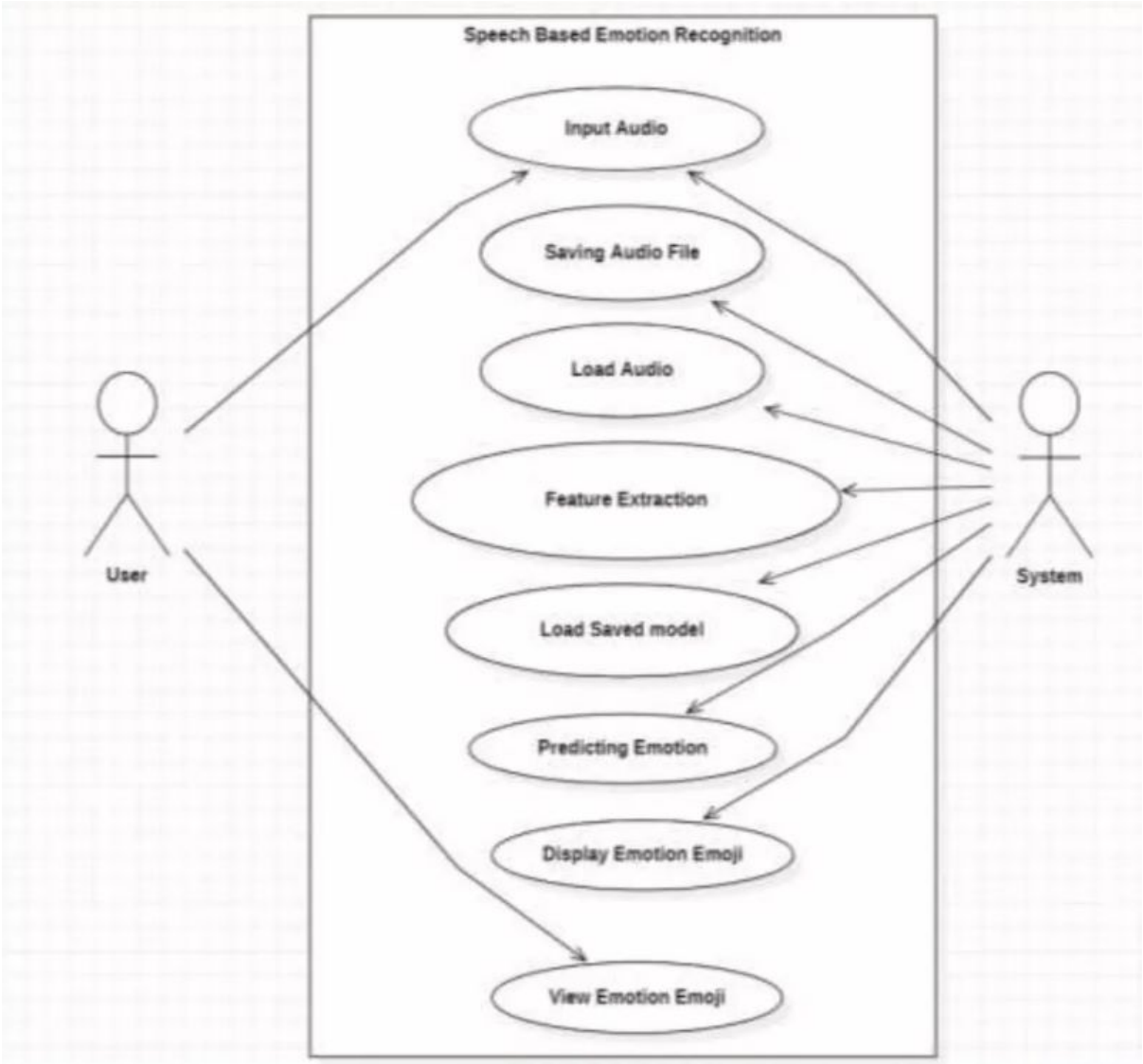


Figure 4.1: Use Case diagram

4.2.2 DFD

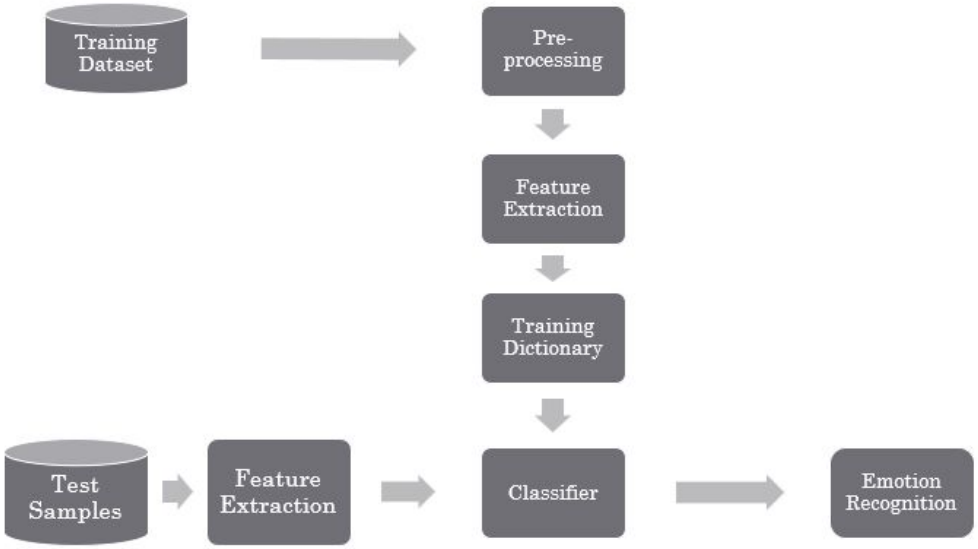


Figure 4.2: DFD

4.2.3 Sequence Diagram

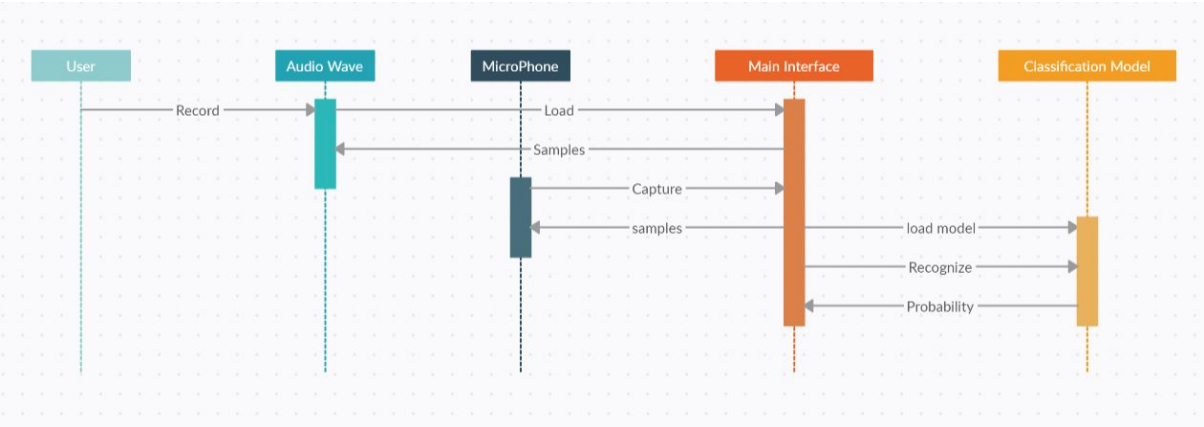


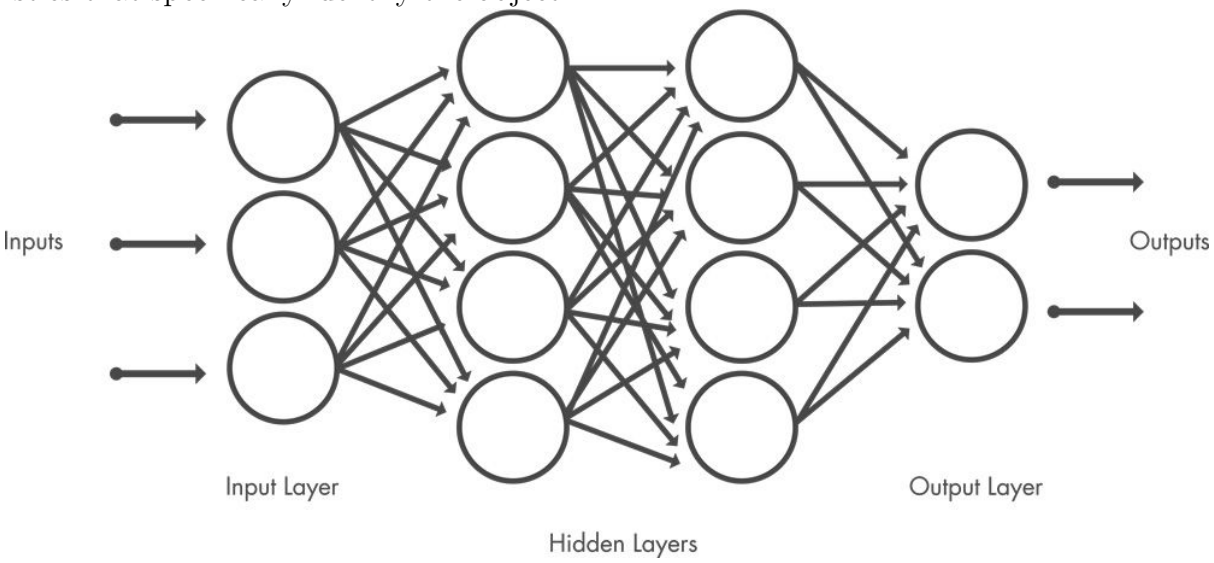
Figure 4.3: Sequence Diagram

4.3 Algorithm and Methodologies

4.3.1 CNN:

A deep learning network architecture known as a convolutional neural network (CNN or ConvNet) learns directly from data, doing away with the requirement for human feature extraction. CNNs are very helpful for recognising objects, faces, and scenes in photos by looking for patterns in the images. For categorising non-image data, such as audio, time series, and signal data, they can be highly useful.

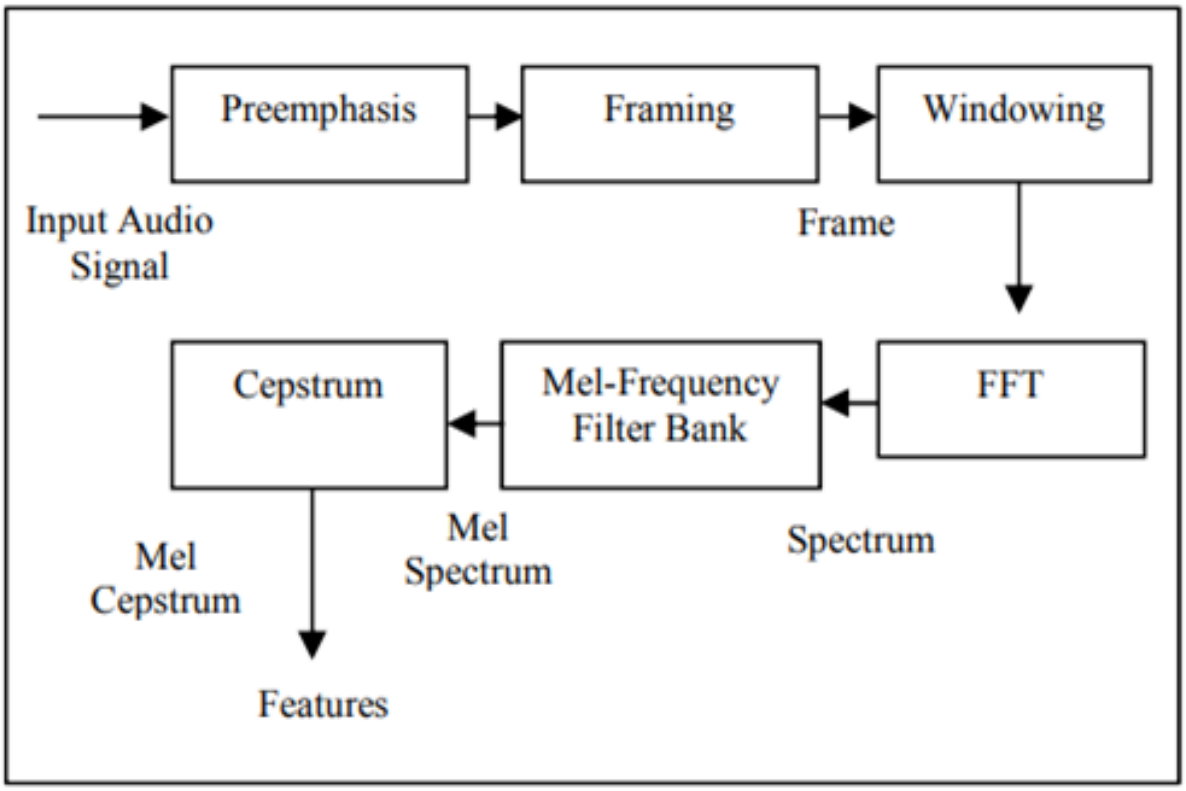
Tens or even hundreds of layers can be present in a convolutional neural network, and each layer can be trained to recognise various aspects of an image. Each training image is subjected to filters at various resolutions, and the result of each convolved image is utilised as the input to the following layer. Beginning with relatively basic properties like brightness and borders, the filters can get more complicated until they reach characteristics that specifically identify the object.



A CNN is made up of an input layer, an output layer, and numerous hidden layers in between, similar to other neural networks.
These layers carry out operations on the data in order to discover characteristics unique to the data. Convolution, activation or ReLU, and pooling are three of the most used layers.

4.3.2 MFCC

MFCC reduces the frequency information of speech signal into the small number of coefficients which is easy to compute and extract the features Pre-emphasis is the process of the Signal with a pre-emphasis filter applied for a smoother spectral appearance. In the frame blocking procedure, the sound signal is divided into a number of short overlapped frames with a frame size of 20 ms and a step of 20 ms between each frame. For the analysis of a segment of long signals, windowing is a necessary step. 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. 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)

5. Implementation

5.1 Stages of Implementation

5.1.1 Data Preprocessing

Any type of processing done on raw data to get it ready for another data processing operation is referred to as data preprocessing, which is a part of data preparation. It has historically been a crucial first stage in the data mining process.

5.1.2 Implementation of Modules

Reading Audio fill

Reading in Audio Files

There are many types of audio files: `mp3` , `wav` , `m4a` , `flac` , `ogg`

In [2]:
```
audio_files = glob('../input/ravdess-emotional-speech-audio/*/*.wav')
```

In [3]:
```
# Play audio file
ipd.Audio(audio_files[0])
```

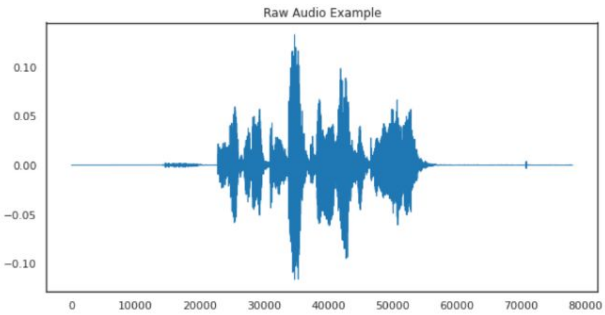
Out[3]: ▶ 0:01 / 0:03

Wave Representation

In [4]:
```
y, sr = librosa.load(audio_files[0])
print(f'y: {y[:10]}')
print(f'shape y: {y.shape}')
print(f'sr: {sr}')
```

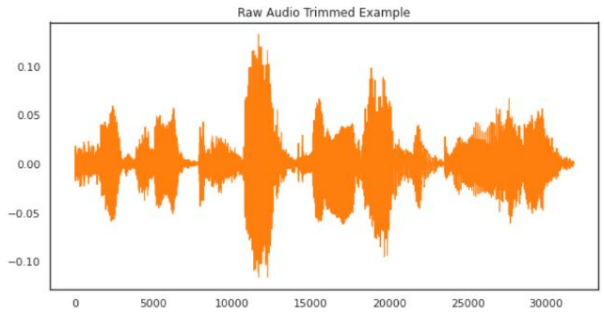
```
y: [0. 0. 0. 0. 0. 0. 0. 0. 0. 0.]
shape y: (77988,)
sr: 22050
```

In [5]:
```
pd.Series(y).plot(figsize=(10, 5),
                  lw=1,
                  title='Raw Audio Example',
                 color=color_pal[0])
plt.show()
```



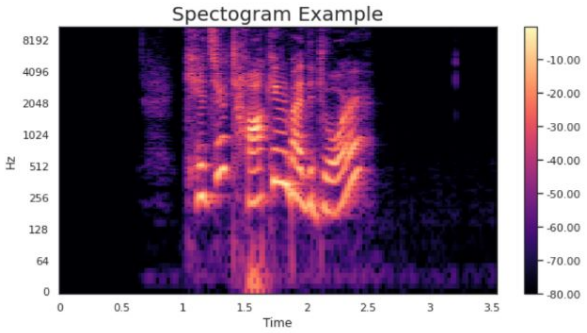
Trimmed audio section

In [6]:
```
# Trimming leading/lagging silence
y_trimmed, _ = librosa.effects.trim(y, top_db=20)
pd.Series(y_trimmed).plot(figsize=(10, 5),
                  lw=1,
                  title='Raw Audio Trimmed Example',
                 color=color_pal[1])
plt.show()
```



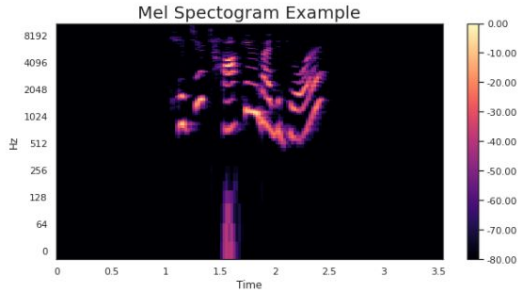
Spectogram

In [9]:
```
# Plot the transformed audio data
fig, ax = plt.subplots(figsize=(10, 5))
img = librosa.display.specshow(S_db,
                               x_axis='time',
                               y_axis='log',
                               ax=ax)
ax.set_title('Spectogram Example', fontsize=20)
fig.colorbar(img, ax=ax, format=f'%0.2f')
plt.show()
```



Mel Spectogram

[17]:
```
fig, ax = plt.subplots(figsize=(10, 5))
# Plot the mel spectogram
img = librosa.display.specshow(S_db_mel,
                               x_axis='time',
                               y_axis='log',
                               ax=ax)
ax.set_title('Mel Spectogram Example', fontsize=20)
fig.colorbar(img, ax=ax, format=f'%0.2f')
plt.show()
```



6. Results

6.1 Results of Experiments

Traing dataset for Speech Emotion Recognition (RAVDESS dataset) is successfully loaded and data preparation have been carried out. Audio samples from the dataset have been visualised using matplotlib and librosa library. Wave form, Spectogram and Mel Spectogram have been successfully visualised for a audio sample

6.2 Conclusion

A speech emotion recognition system (SER) utilising the CNN model and MFCC feature extraction approach is described in this study as an enhanced mechanism for SER. The input to deep CNNs is a spectrogram representing the spoken signal. It has been noted that MFCC is a frequently used feature and improves SER's ability to identify emotions. Each spectrogram that is created for the incoming voice sounds is predicted using the trained model.

By processing only the segments chosen for emotion recognition rather than all segments complying with a computational social system, we shorten the processing time of our system. We could tell whether the speaker was a man or a woman by their gender. However, there is room for improvement by combining several features and refining the ML model for a higher true positive rate.

6.3 Limitations of the Project

It is a challenge to make emotion available in different languages.

There are limitations with different types and versions of the software such as dataset input are only textual data, and image, pattern, video, and audio inputs are invalid.

Each emotion may correspond to the different portions of the spoken utterance. The same utterance may show different emotions Therefore it is very difficult to differentiate these portions of utterances. Another problem is that the Expression of emotion is depending on the speaker and their culture and environment.

6.4 Future Scope

• The suggested architecture can be used in the future for additional applications and can be used to improve accuracy and reduce computational complexity for voice emotion recognition using GRU, DBN, and spike networks.

• It will also be utilised to determine whether the speaker is a man or a woman. will be more accurately and perfectly able to identify emotions like happiness, sadness,

disgust, anger, etc.

• The suggested model can be a goal for speaker identification and recognition that is applied to numerous real-world issues.

• We are developing a web application for clinical usage that can receive audio input and predict the emotion. The application will be user-interactive, allowing the user to input speech and determine the emotion of that input.

Bibliography

[1] B. Schuller, G. Rigoll, and M. Lang, "Hidden markov model-based speech emotion recognition," in 2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings. (ICASSP'03)., vol. 2. IEEE, 2003, pp. II–1

[2] K. Han, D. Yu, and I. Tashev, "Speech emotion recognition using deep neural network and extreme learning machine," in Fifteenth annual conference of the international speech communication association, 2014

[3] S. K. Bhakre and A. Bang, "Emotion recognition on the basis of audio signal using Naive Bayes classifier," 2016 International Conference on Advances in Computing, Communications and Informatics (ICACCI), 2016, pp. 2363-2367, doi: 10.1109/ICACCI.2016.7732408.

[4] S. Mirsamadi, E. Barsoum and C. Zhang, "Automatic speech emotion recognition using recurrent neural networks with local attention," 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2017, pp. 2227-2231, doi: 10.1109/ICASSP.2017.7952552.

[5] Noroozi, F., Sapiński, T., Kamińska, D. et al. Vocal-based emotion recognition using random forests and decision tree. Int J Speech Technol 20, 239–246 (2017). https://doi.org/10.1007/s10772-017-9396-2

[6] A. M. Badshah, J. Ahmad, N. Rahim and S. W. Baik, "Speech Emotion Recognition from Spectrograms with Deep Convolutional Neural Network," 2017 International Conference on Platform Technology and Service (PlatCon), 2017, pp. 1-5, doi: 10.1109/PlatCon.2017.7883728.

[7] Zhang Y YDu JWang Z Ret al. Attention based recognition [C]2018 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference. Honolulu, USA: IEEE, 2018: 1771-1775

[8] Mustaqeem, M. Sajjad and S. Kwon, "Clustering-Based Speech Emotion Recognition by Incorporating Learned Features and Deep BiLSTM," in IEEE Access, vol. 8, pp. 79861-79875, 2020, doi: 10.1109/ACCESS.2020.2990405.

[9] T. M. Wani, T. S. Gunawan, S. A. A. Qadri, H. Mansor, M. Kartiwi and N. Ismail, "Speech Emotion Recognition using Convolution Neural Networks and Deep Stride Convolutional Neural Networks," 2020 6th International Conference on Wireless and Telematics (ICWT), 2020, pp. 1-6, doi: 10.1109/ICWT50448.2020.9243622.

Sample Document

Sample Document

Sample Document

Sample Document