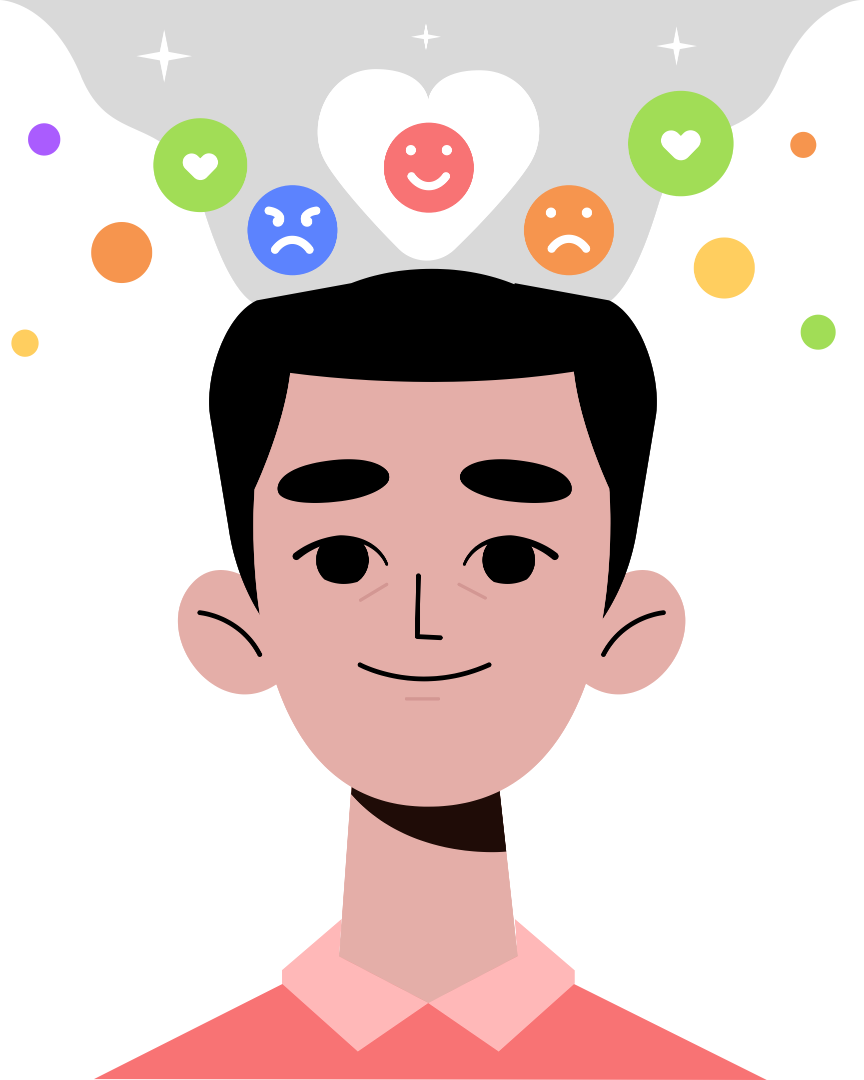


Speech Emotion Recognition

Group 2

Aneshaa Kasula - 014558427

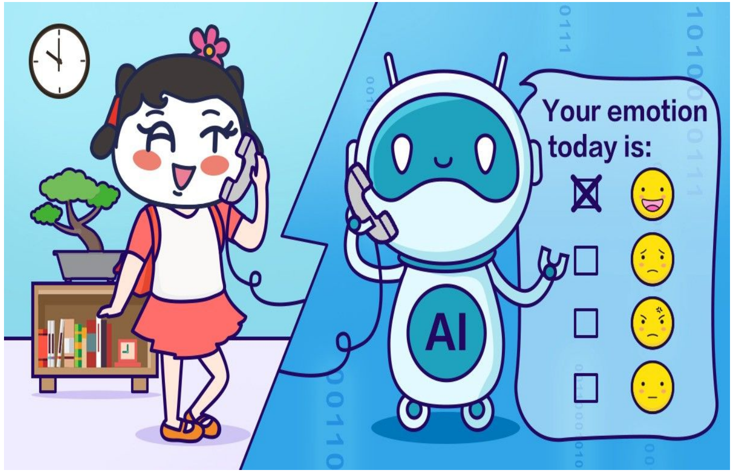
Bailey Wang - 010461113

Rudra Gandhi - 011066640

Viritha Vanama - 015356991

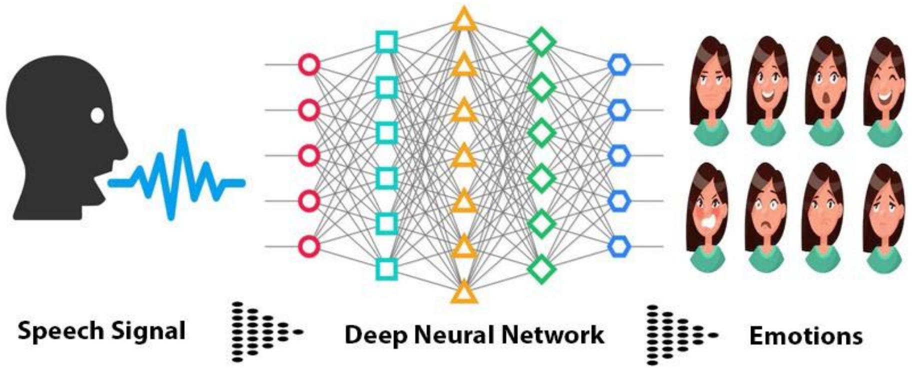
Introduction

- ❑ Speech Emotion Recognition (SER) recognize emotional aspect of speech
- ❑ Humans are naturally able to identify tones and can determine the emotion of the speaker
- ❑ Understanding the emotion can be useful in real-time situations like phone calls, call center operators, or customer service
- ❑ Using Deep Learning Models, and eventually have the AI provide an appropriate response to situations



Objective & Scope

- ❑ Recognize the emotion from the audio input
- ❑ Build a Neural Network to identify and classify emotion
- ❑ Real-time Emotion Detection using the tone of their voice
- ❑ Restrictive to English language
- ❑ American accent



Technology Survey

Model	Pros	Cons
Multilayer Perceptron (MLP)	-Can be applied to complex non-linear problems -Works well with large data input -Provides quick prediction after training	-Computations are difficult and time consuming -Proper functioning of the model depends on the quality of training
Convolution Neural Network (CNN)	-Speed -Great for short texts	-Does not have a sense of memory state, and they are rather limited for sequential modelling (such as language models, speech recognition etc.)
Long-Short Term Memory (LSTM)	-Provide us with a large range of parameters such as learning rates, and input and output biases. Hence, no need for fine adjustments.	-LSTMs are prone to overfitting and it is difficult to apply the dropout algorithm to curb this issue.

Literature Survey

Paper	Objective	Methodology Used	Result
P. Harár, R. Burget and M. K. Dutta, "Speech emotion recognition with deep learning," *2017*	This paper describes a method for Speech Emotion Recognition using Deep Neural Network with convolutional, pooling, and fully connected layers.	-Stochastic Gradient Descent	Overall accuracy was 79.14%
Linqin Cai, Yaxin Hu, Jiangong Dong, Sitong Zhou, "Audio-Textual Emotion Recognition Based on Improved Neural Networks",	In this paper, they introduce two different models in their speech recognition system	-CNN -LSTM	Overall accuracy was 75%
Trinh Van, L.; Dao Thi Le, T.; Le Xuan, T.; Castelli, E. Emotional Speech Recognition Using Deep Neural Networks	In this paper, they are trying to make a connection between vocal emotion and physical emotion	-CNN -GRU	Average accuracy was 95%

Project Methodology (CRISP - DM)

1 Business and Data Understanding

Create a real-time Neural Network Model to identify human emotion

Audio Data Collection Process

Audio Data Usage in Real Work Scenarios

Audio Data and Metadata understanding

2/21 - 3/09

2 Data Exploration and Data Preparation

Audio Features: Samples, Sample rate, Channels, Bit Depth

Data Cleaning

Data Distribution

03/10 - 04/03

3 Model Implementation

Deep learning Model Implementation for Emotion Detection

Training and testing Neural Network model

Tuning and improving Neural Network model

04/04 - 04/25

4 Model Evaluation

Comparing Neural Network model performance score

Choose the best Neural Network model

04/26 - 05/09

5 Deployment

Final Project Report

Create Prototype

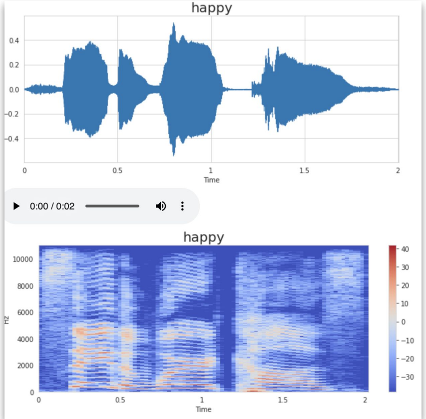
05/10 - 05/16

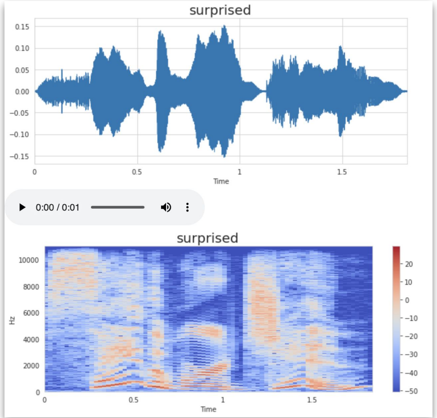
About Data

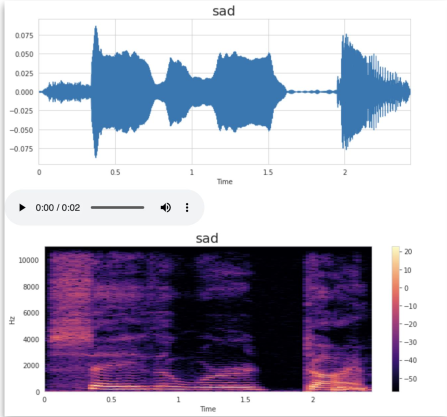
- Using 2 datasets from Kaggle
  - Ravdess
    - Total 1440 Files, 60 files per actor (24 actors)
    - 12 female and 12 male actors (no different age groups)
    - North American Accent
    - Language: English
    - Consists of emotions like: sad, happy, angry, fear, surprise,disgust, neutral
    - Each file ranges for 3 seconds to 5 seconds
- Tess
  - Total of 2800 Files, 200 files per emotion (14 emotions)
  - Only two female actors (age 26 and age 64)
  - North American Accent
  - Language: English
  - Consists of emotions like: sad, happy, angry, disgust, fearful, calm, pleasant_surprise
  - Each file ranges for 3 seconds to 5 seconds

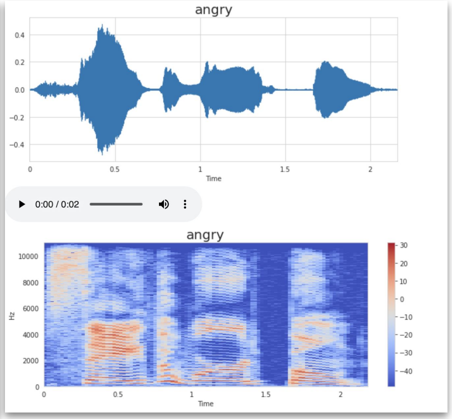
	speech	label
0	/content/drive/Shareddrives/DATA255/Tess/OAF_F...	fear
1	/content/drive/Shareddrives/DATA255/Tess/OAF_F...	fear
2	/content/drive/Shareddrives/DATA255/Tess/OAF_F...	fear
3	/content/drive/Shareddrives/DATA255/Tess/OAF_F...	fear
4	/content/drive/Shareddrives/DATA255/Tess/OAF_F...	fear

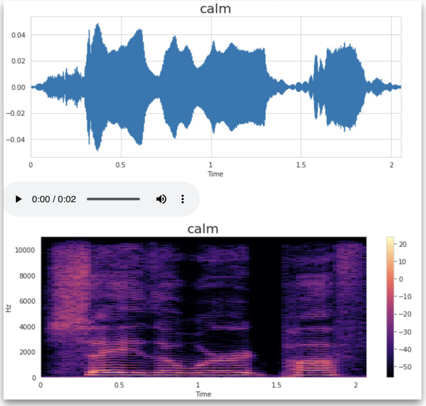
Exploratory Data Analysis

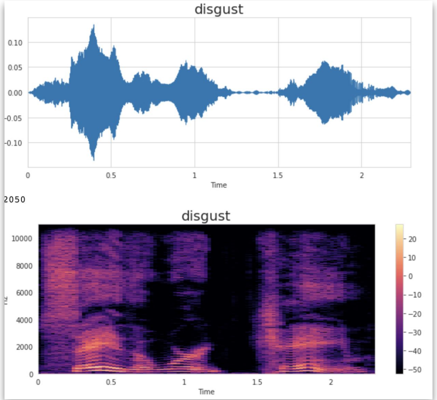


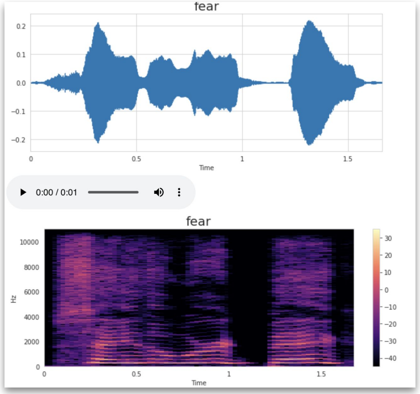








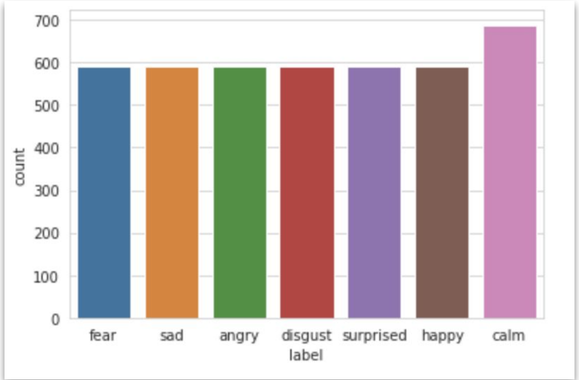




Data Preprocess

Data Transformation:

- ❑ Combine both the datasets,
  - ❑ Combine same emotions with different labels like ‘fear’ and ‘fearful’
  - ❑ Count of all emotions are almost equal
  - ❑ There are 4,240 records of data
- ❑ Data Augmentation:
  - ❑ To increase the dataset size
  - ❑ Added noise to the data
  - ❑ Added pitch to the data to change the pitch
  - ❑ Around 12,000 records

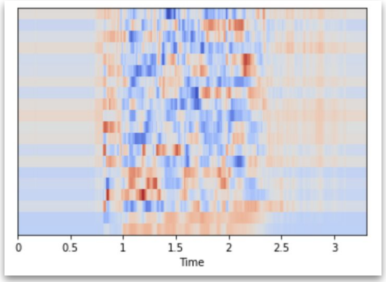



```python
def noise(filename):
    aug = naa.NoiseAug()

    y, sr = librosa.load(filename, duration=3, offset=0.5)

    augmented_data = aug.augment(y)

    mfcc = np.mean(librosa.feature.mfcc(y=augmented_data, sr=sr, n_mfcc=40).T, axis=0)
    return mfcc
```

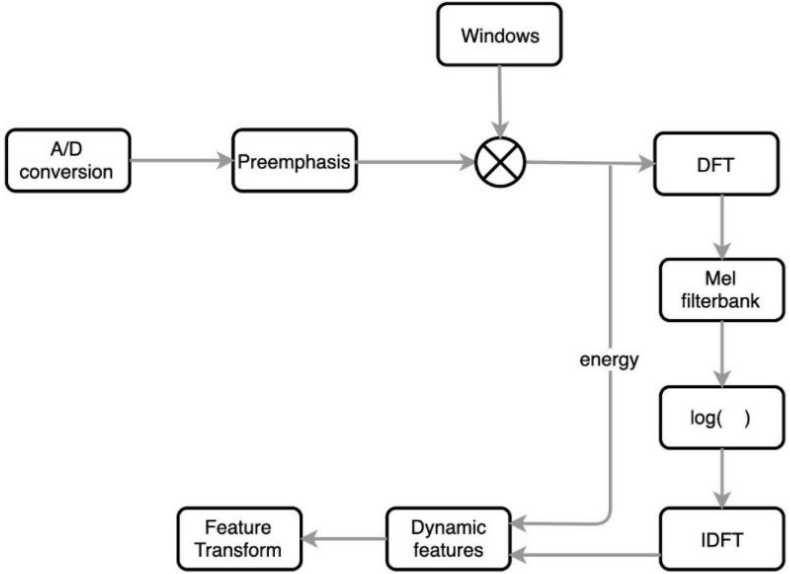
```python
def pitch(filename):

    y, sr = librosa.load(filename, duration=3, offset=0.5)

    aug = naa.PitchAug(sampling_rate=sr, factor=(2,3))
    augmented_data = aug.augment(y)
    mfcc = np.mean(librosa.feature.mfcc(y=augmented_data, sr=sr, n_mfcc=40).T, axis=0)
    return mfcc
```

Feature Extraction

- Spectrograms -> Frequency Vs Time and Amplitude indicated by color
- Mel Frequency Cepstral Coefficients (MFCC) essentially take Mel Spectrograms and apply a couple of further processing steps.
- The inverse of the log of the magnitude of the signal is called a cepstrum.
- It extracts a much smaller set of features from the audio that are the most relevant in capturing the essential quality of the sound.

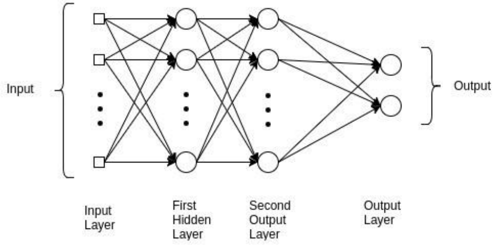


<class 'numpy.ndarray'> (20, 310)

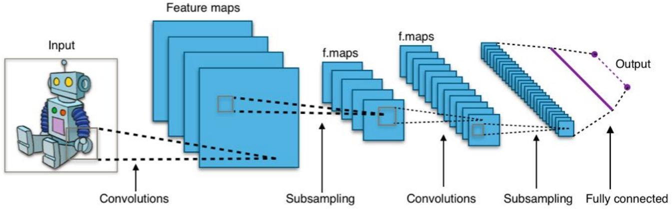


Model Selection

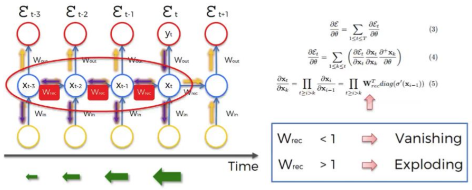
- ❑ **Multilayer Perceptron (MLP)**



- ❑ **Convolution Neural Network (CNN)**



- ❑ **Long-Short Term Memory (LSTM)**



Model Selection Continued...

Model	Why we chose?	How we applied the model?
Multilayer Perceptron (MLP)	Baseline point of comparison Flexible working with different types of data inputs, maps inputs to outputs	MLP with input layer, 8 hidden layers (funnel neurons), learning rate = .003
Convolution Neural Network (CNN)	Able to handle 2D input, often used when there is a ordered relationship (time)	CNN with input, convolution and maxpooling layers, dropout once the Convolution/maxpooling is completed, funnel dense layers
Long-Short Term Memory (LSTM)	The memory concept of the LSTM is useful in capturing Long-term patterns in audio data	LSTM with input and Dense layers with Batch Normalization and Dropout layer

Model Summary

MLP

```
mlp=MLPClassifier(solver='adam', batch_size=64,
                  hidden_layer_sizes= (400, 350, 300, 250, 200, 150, 100, 50),
                  activation = 'relu',
                  learning_rate='adaptive',
                  learning_rate_init=0.003,
                  max_iter=500,
                  verbose = 1)
```

LSTM

Layer (type)	Output Shape	Param #
lstm_9 (LSTM)	(None, 512)	1052672
batch_normalization_9 (BatchNormalization)	(None, 512)	2048
dense_18 (Dense)	(None, 64)	32832
dropout_11 (Dropout)	(None, 64)	0
dense_19 (Dense)	(None, 7)	455

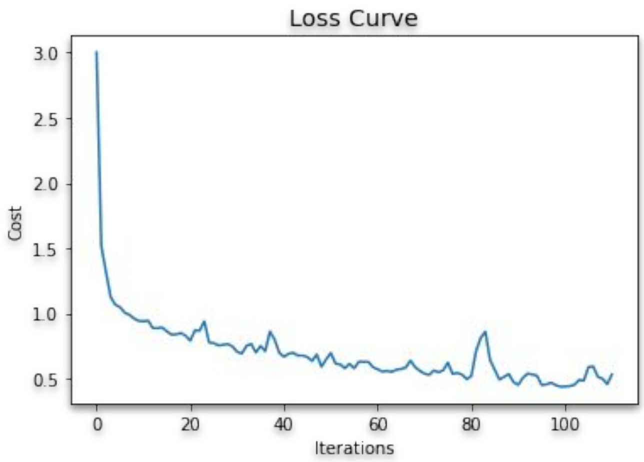
Total params: 1,088,007
Trainable params: 1,086,983
Non-trainable params: 1,024

CNN

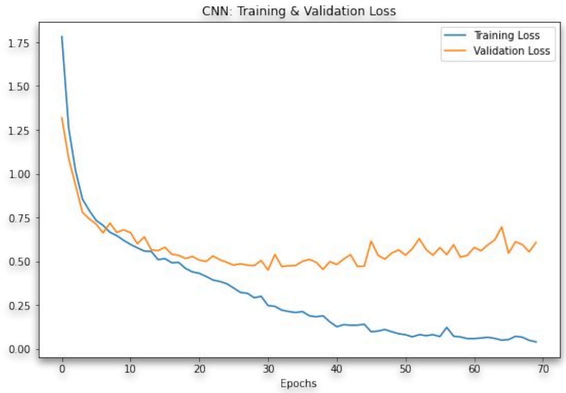
Layer (type)	Output Shape	Param #
conv1d (Conv1D)	(None, 40, 64)	576
zero_padding1d (ZeroPadding1D)	(None, 42, 64)	0
max_pooling1d (MaxPooling1D)	(None, 42, 64)	0
conv1d_1 (Conv1D)	(None, 42, 64)	32832
zero_padding1d_1 (ZeroPadding1D)	(None, 44, 64)	0
max_pooling1d_1 (MaxPooling1D)	(None, 44, 64)	0
conv1d_2 (Conv1D)	(None, 44, 128)	65664
zero_padding1d_2 (ZeroPadding1D)	(None, 46, 128)	0
max_pooling1d_2 (MaxPooling1D)	(None, 46, 128)	0
conv1d_3 (Conv1D)	(None, 46, 128)	131200
zero_padding1d_3 (ZeroPadding1D)	(None, 48, 128)	0
max_pooling1d_3 (MaxPooling1D)	(None, 48, 128)	0
dropout (Dropout)	(None, 48, 128)	0
flatten (Flatten)	(None, 6144)	0
dense (Dense)	(None, 256)	1573120
dropout_1 (Dropout)	(None, 256)	0
dense_1 (Dense)	(None, 128)	32896
dropout_2 (Dropout)	(None, 128)	0
dense_2 (Dense)	(None, 64)	8256
dropout_3 (Dropout)	(None, 64)	0
dense_3 (Dense)	(None, 32)	2080
batch_normalization (BatchNormalization)	(None, 32)	128
dense_4 (Dense)	(None, 7)	231

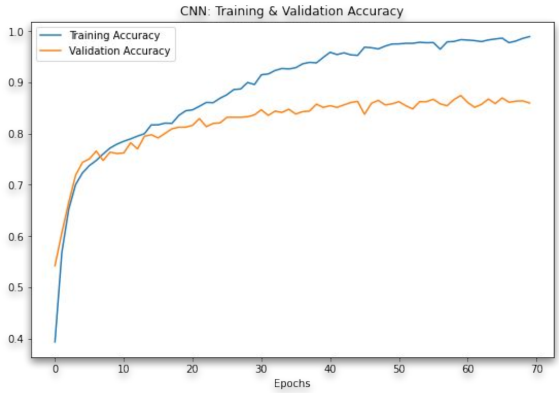
Model Evaluation

MLP

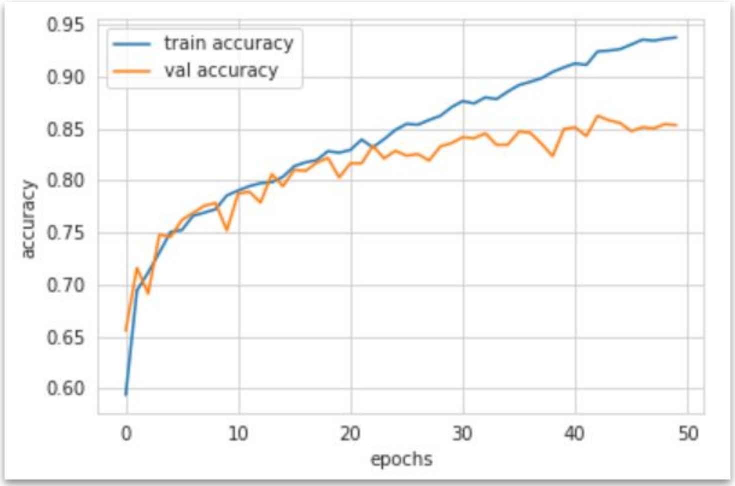


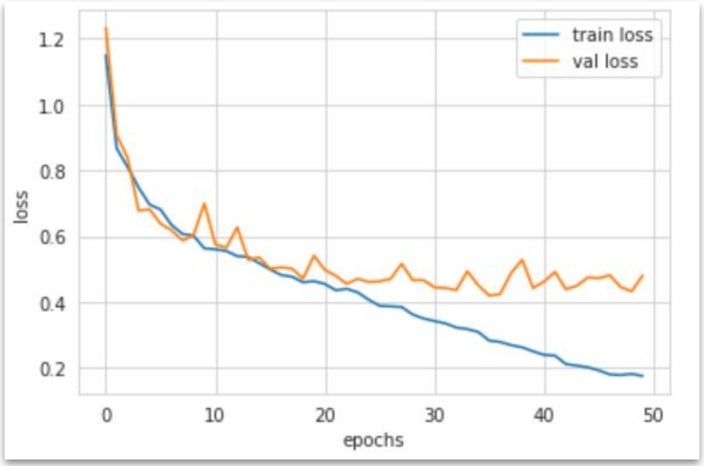
CNN





Model Evaluation - LSTM





Model Evaluation Continued..

Model	Insights	Accuracy
MLP	-Simple to implement, create hidden layer and add number of neurons -custom learning rate -Often low results -Unable to add Batch_Normalization or Dropout	79%
CNN	-Batch_Normalization often worsened the accuracy -Funnel method took a long time to compute -overfitting -Handling bias and variance was difficult	84%
LSTM	-LSTM is the best approach for audio data when compared to MLP and CNN. -Overfitting is a real issue, handled using Dropout, custom learning rate -Handling bias and variance was difficult; Adamax handled the variance	85.4%

Application

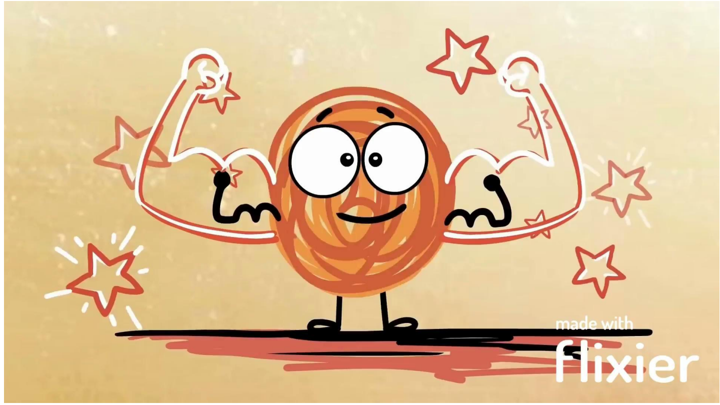
- ❏ Voice Recognition is becoming more and more popular
- ❏ Useful in monitoring a person’s psychological state
- ❏ Speech recognition can be used in marketing, healthcare, customer satisfaction, gaming experience, stress monitoring

Challenges & Future Scope

- Audio is one of the challenging datasets
- Very few open-source Audio Datasets available
- Domain knowledge in Signal processing is required
- Preprocessing steps has very few references
- Feature extraction is really slow, so used ‘Swifter’ module
- High Computation power is required for feature extraction

- We can train our models on different languages and different accents
- Extend the models to classify male and female voices
- Identify the individual person's voice like google mini

Demo



Questions

References

- https://alibabatech.medium.com/voice-based-emotion-recognition-framework-for-films-and-tv-programs-2a6abbb77242
- https://github.com/SuyashMore/MevonAI-Speech-Emotion-Recognition
- https://www.analyticsvidhya.com/blog/2020/12/mlp-multilayer-perceptron-simple-overview/
- https://en.wikipedia.org/wiki/Convolutional_neural_network
- https://www.superdatascience.com/blogs/recurrent-neural-networks-rnn-long-short-term-memory-lstm
- https://www.analyticsvidhya.com/blog/2021/06/mfcc-technique-for-speech-recognition/
- https://ieeexplore.ieee.org/abstract/document/8049931
- https://www.hindawi.com/journals/mpe/2019/2593036/
- https://www.mdpi.com/2227-7390/8/12/2133/htm
- https://ieeexplore.ieee.org/abstract/document/7952552
- https://www.mdpi.com/1424-8220/22/4/1414/htm
- https://towardsdatascience.com/multilayer-perceptron-explained-with-a-real-life-example-and-python-code-sentiment-analysis-cb408ee93141
- https://towardsdatascience.com/lstm-networks-a-detailed-explanation-8fae6aefc7f9
- https://towardsdatascience.com/convolutional-neural-networks-explained-9cc5188c4939
- https://github.com/makcedward/nlpaug/blob/master/example/audio_augmenter.ipynb