



Audio Emotion Classification

Ridzuan Mokthar





Presentation Overview

01//

**Problem statement
Background &
Datasets**

//02

**Preprocess
& Feature Extraction**

03//

**Model Architecture,
Model Evaluation &
Application**

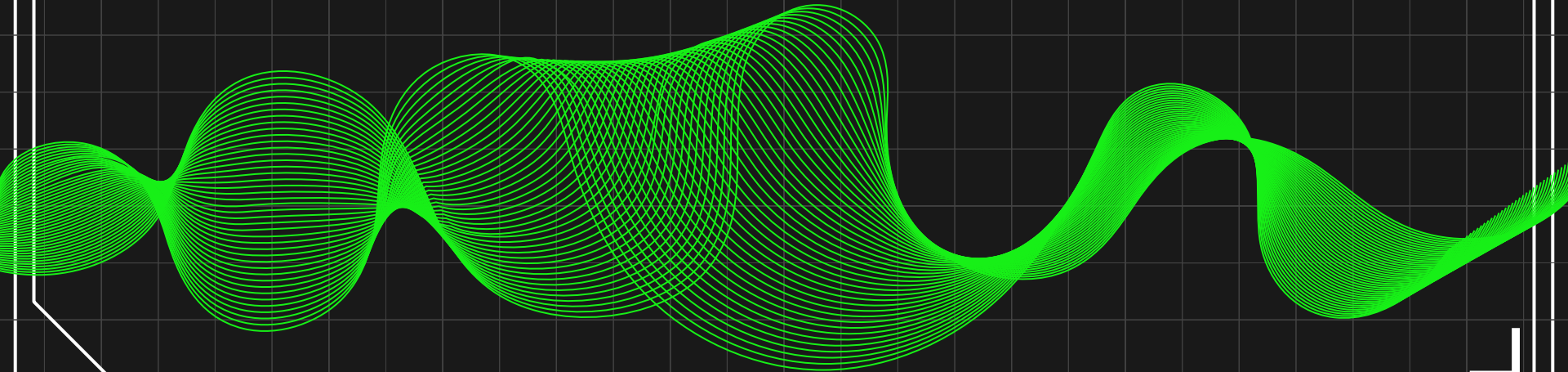
//04

**Conclusion
& Recommendation**



01//

Background & Problem statement





Background

- Communication plays a vital role in building and maintaining relationships with one another.
- Emotional projection is one of the first telltale signs that we notice or observe during a conversation.
- Our emotions can not only be portrayed by what words we say but also by how we say them.



Identifying the Problem



Emotional quotient, Empathy and Alexithymia

- Low Emotional quotient (EQ) is the inability to distinguish emotions in both self and others.
- Cognitive Empathy is the lack of understanding for how others feel.
- Alexithymia is when a person has difficulty identifying and expressing emotions.

Why should we solve it



Foreseen inconveniences

- Poor performance at school or work from lack of communication, inability to express or understand their grievances.
- Degrading Physical & Mental well-being due to stress in turn may lead to a more serious situation such as depression or suicide.
- Failure to foster relationships may lead to them being dysfunctional.
- Low Social Intelligence due to absence of experience with people.
- Violent tendencies induced by stress or confusion



Solution & Benefits



How we plan to solve it?

- CNN Deep Learning
- Prediction Model to predict emotion from speech
- Application to provide educational aid



What are the benefits?

- Given the ability to identify root emotions.
- Improving and also further enforcing their knowledge related to emotions

Datasets

Toronto emotional speech set (TESS)

- 2800 audio files (WAV format)
- 2 voice actors (Old & Young)
- 7 emotions: anger, disgust, fear, happiness, neutral, surprise & sadness

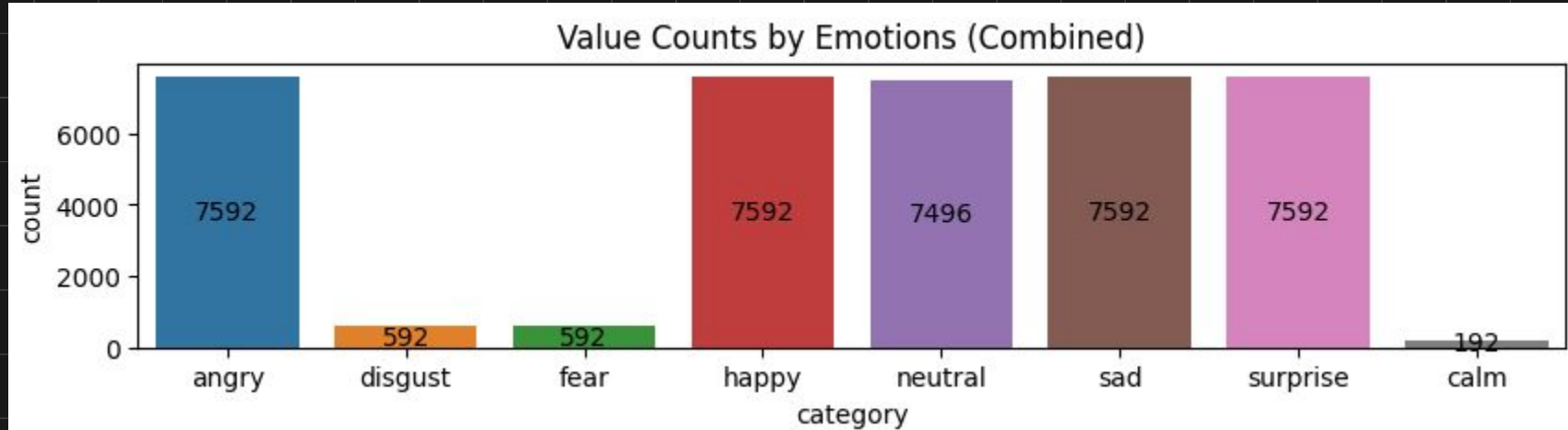
Emotional Speech Dataset (ESD)

- 3500 audio files (WAV format)
- 10 Mandarin & 10 English speakers
- 5 emotions: neutral, happy, angry, sad & surprise

Ryerson Aud-Vis DB of Emo Speech & Song (RAVDESS)

- 1440 audio files (WAV format)
- 24 voice actors (12 male & 12 female)
- 7 emotions: calm, happiness, sadness, anger, fear, disgust and surprise

Datasets (Combined)

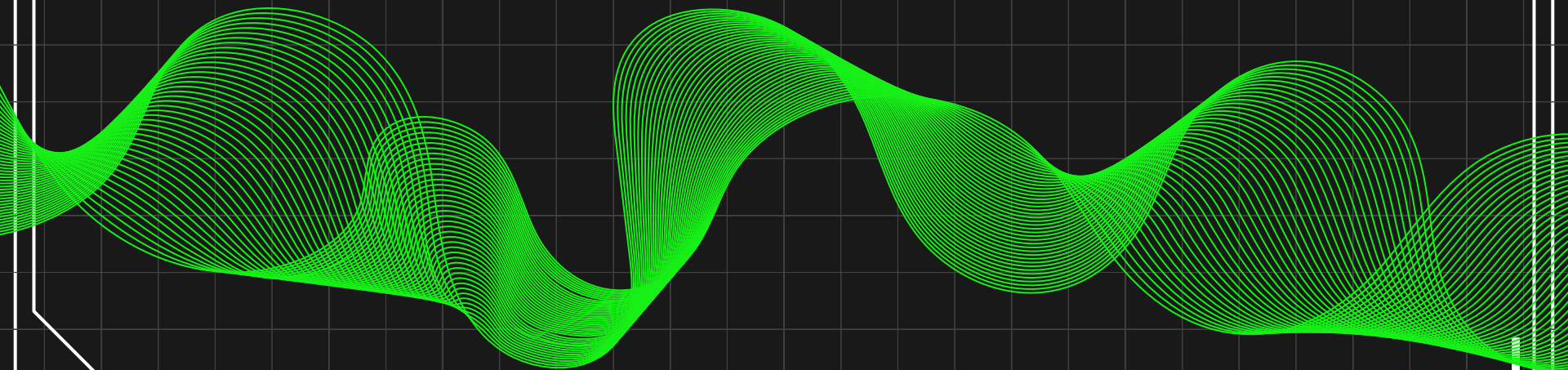


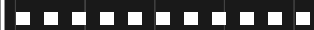
- Disgust, Fear, and Calm will be dropped due to low observation after combining the 3 dataset. Otherwise they will be noise and affect the model prediction capabilities.



02//

Preprocess & Feature Extraction

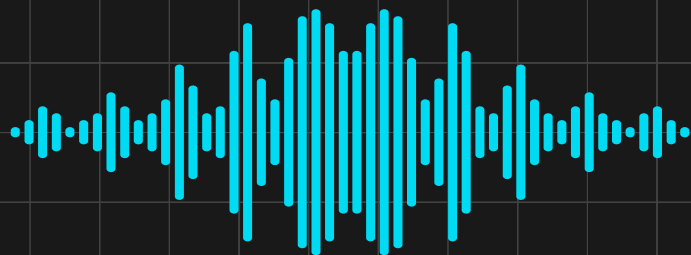




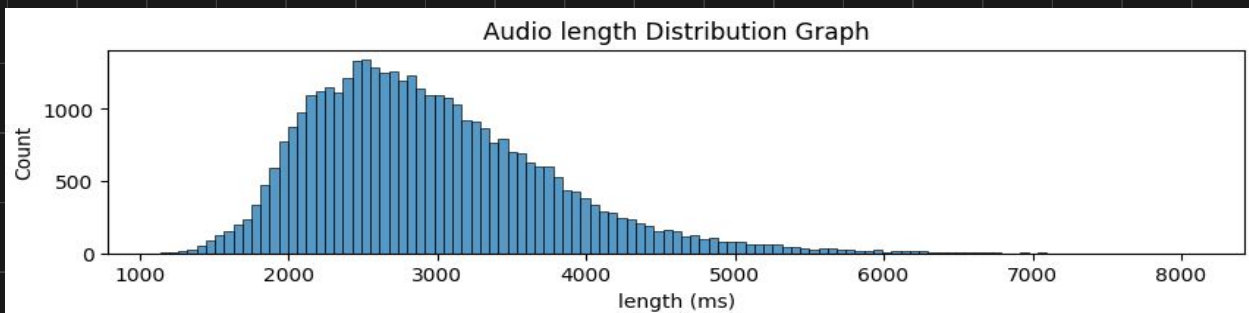
Audio data processing

Standardize basic Audio format/parameters using Librosa:

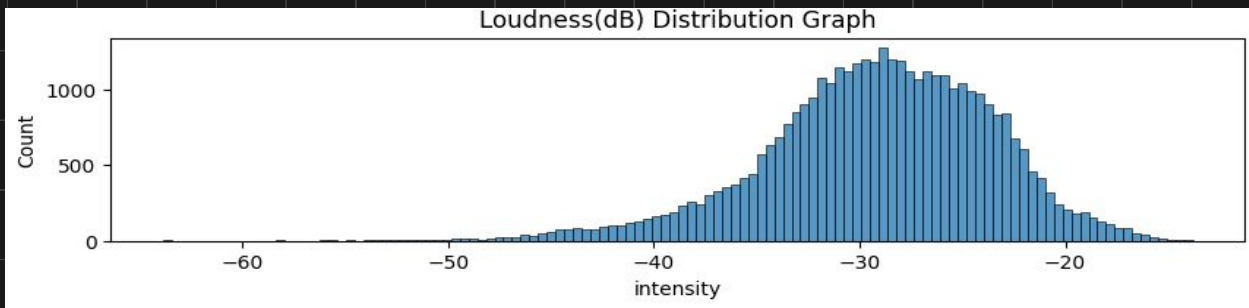
- Channels: 1 for mono
- Bit depth: 2 for 16-bit
- Sample rate: 22050 Hz



Metadata Inference



- Max audio length: 8080 ms
- Min audio length: 1139 ms



- Loudest audio: -14dBFS
- Softest audio: -64dBFS



Audio Feature Extraction

Features extracted and used for modeling:

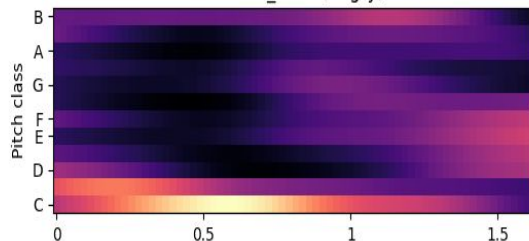
- 1) Chroma energy normalized statistics (CENS)
- 2) Spectral Bandwidth
- 3) Spectral Centroid
- 4) Mel-frequency cepstral coefficients (MFCC)
- 5) Root Mean Square Energy
- 6) Tonal Centroid Features (Tonnetz)



Chroma energy norm statistics

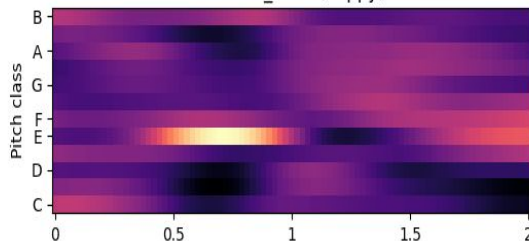
Anger

Chroma_CENS(angry)



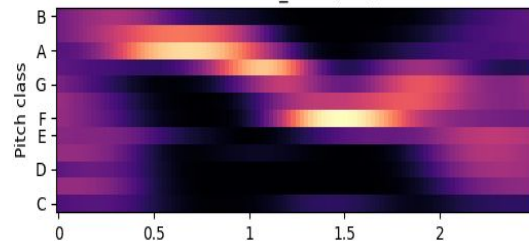
Happiness

Chroma_CENS(happy)



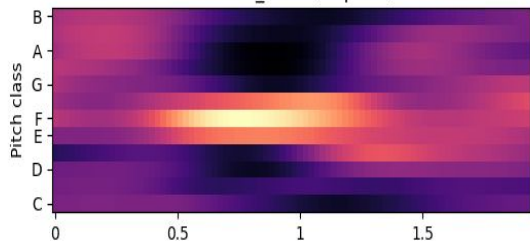
Sadness

Chroma_CENS(sad)



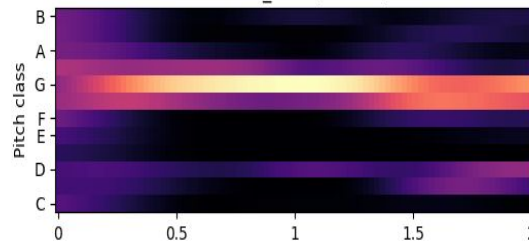
Surprised

Chroma_CENS(surprise)



Neutral

Chroma_CENS(neutral)



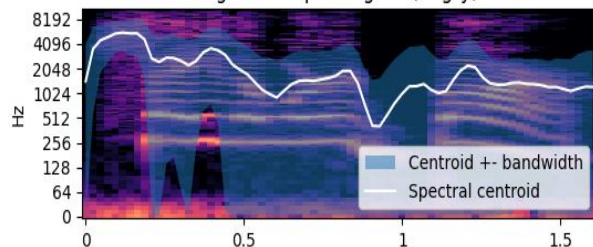
- When using CENS we can see certain pitch class are unique to each emotion.



Spectral Bandwidth + Centroid

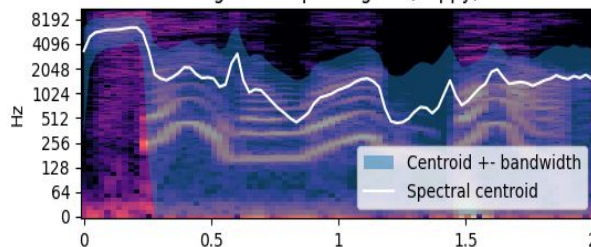
Anger

log Power spectrogram(angry)



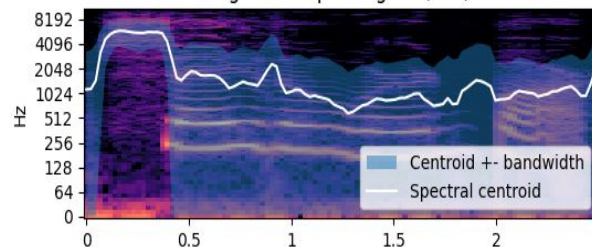
Happiness

log Power spectrogram(happy)



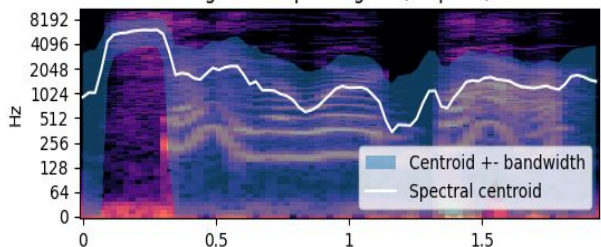
Sadness

log Power spectrogram(sad)



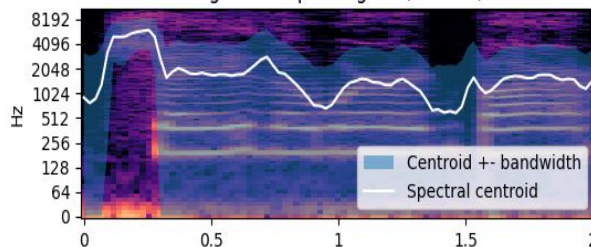
Surprised

log Power spectrogram(surprise)

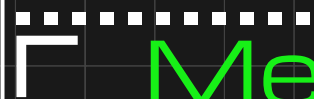


Neutral

log Power spectrogram(neutral)

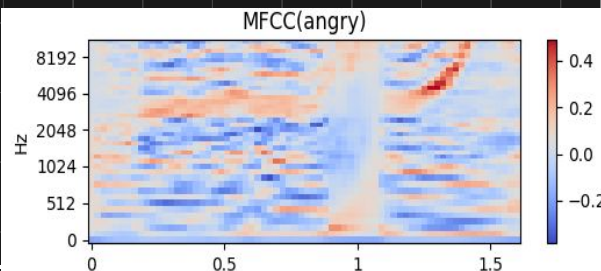


- From this single example we may not be able to visually differentiate between the emotions

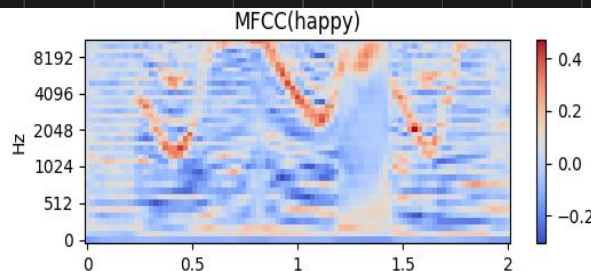


Mel-frequency cepstral coef.

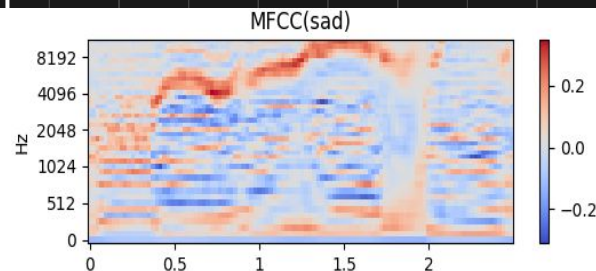
Anger



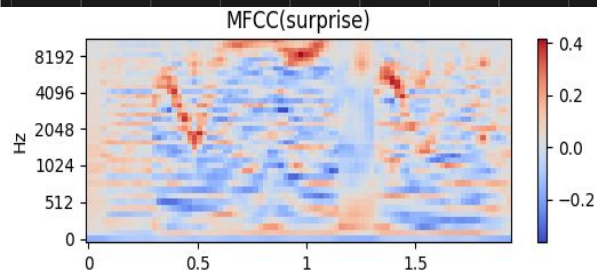
Happiness



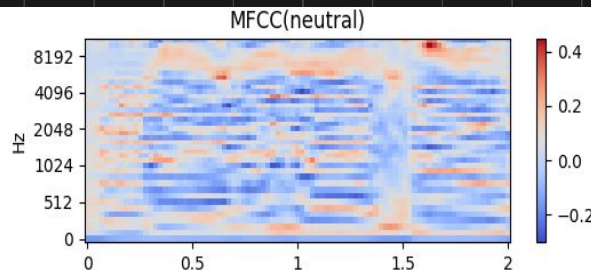
Sadness



Surprised



Neutral



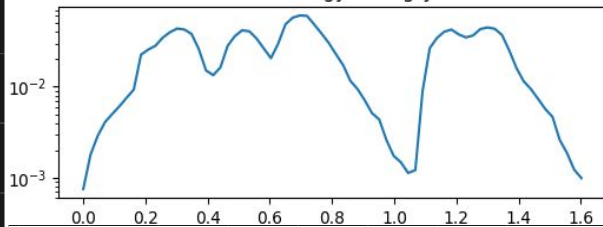
- MFCC holds the entirety of information related to the audio.



Root Mean Square Energy

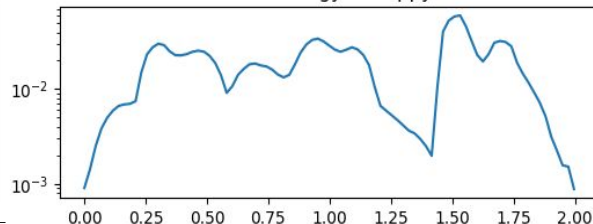
Anger

RMS Energy of angry



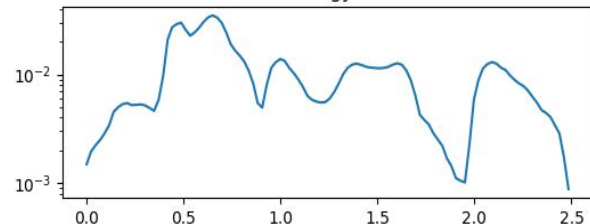
Happiness

RMS Energy of happy



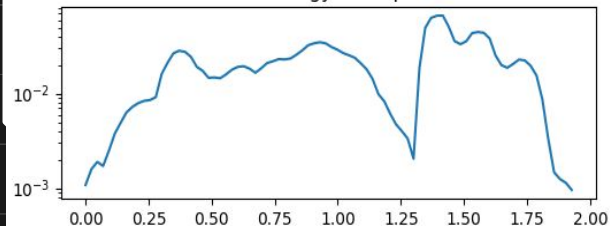
Sadness

RMS Energy of sad



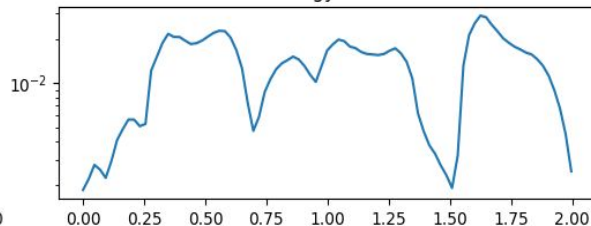
Surprised

RMS Energy of surprise



Neutral

RMS Energy of neutral

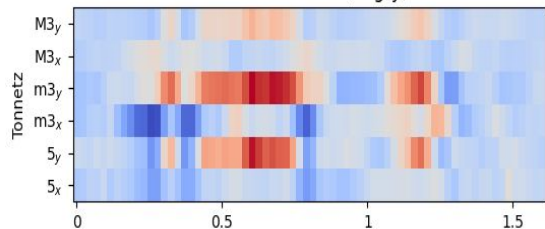


- We can see some minor difference between emotion when observing Peak to valley and also the smoothening overall graph.

Tonal Centroid Features

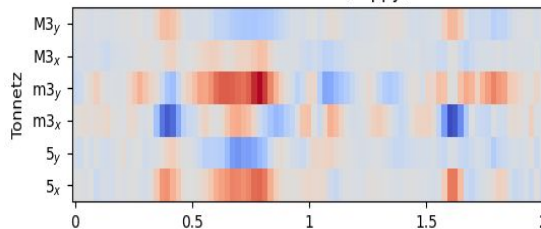
Anger

Tonal Centroids (angry)



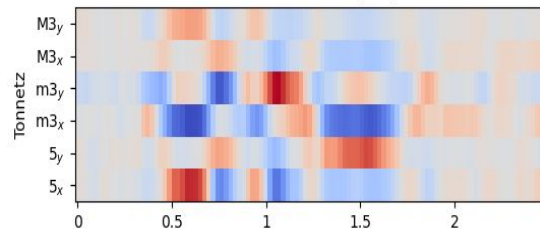
Happiness

Tonal Centroids (happy)



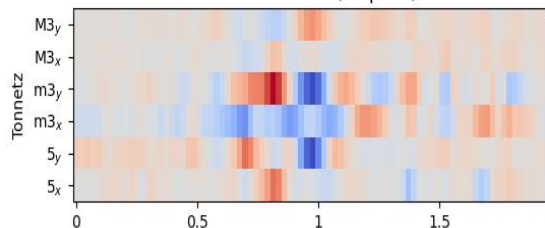
Sadness

Tonal Centroids (sad)



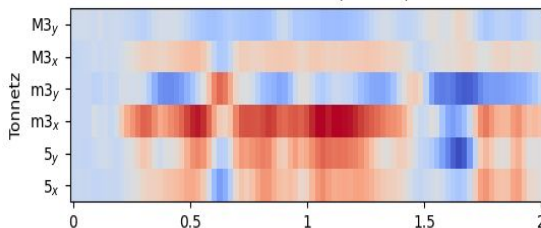
Surprised

Tonal Centroids (surprise)



Neutral

Tonal Centroids (neutral)



- From here we can see tonal centroid features which are unique to the different emotion class



03//

Model Architecture & Evaluation



Model Evaluation

Averaging Techniques for Multiclass classification

- Since our classes are balance we will be mostly looking at Macro average.
- Macro Average : A simple arithmetic mean of all metrics across classes. This technique gives equal weights to all classes making it a good option for balanced classification tasks.

Metrics Scores

- 1) Macro Average Accuracy will be the main scoring metric to assess how well the model predict TP & TN.
- 2) Sub metric will be macro average of Precision & Recall
- 3) Additional metric is Matthew's correlation coefficient, it ranges from -1 to 1 where 0 means the model is no better than random chance.



CNN Deep Learning Model

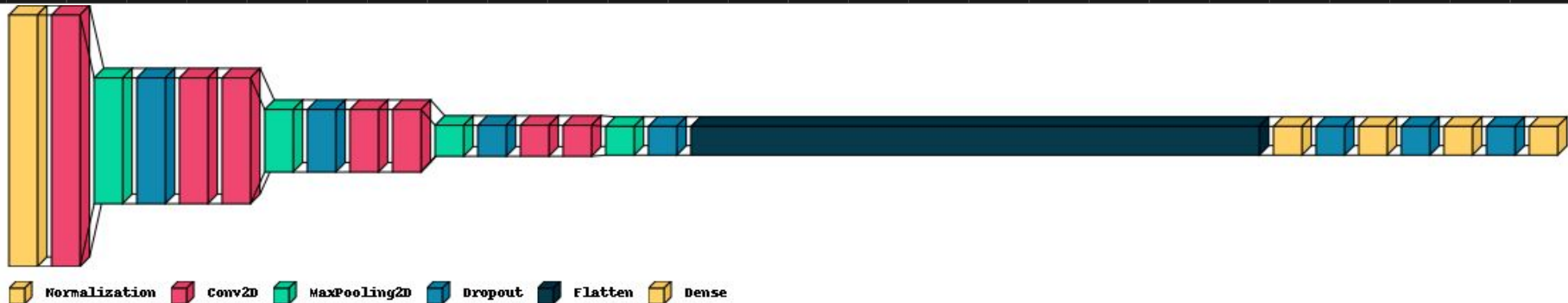
Why CNN Deep Learning was chosen?

- The Convolutional Neural Network built-in convolutional layer reduces the high dimensionality of images without losing its information.

Each Audio feature is a “grayscale image”

- For n features will be our n dimension of the “image” and will undergo convolution & max pooling.
 - The final output of the convolutional layer is an array vector of max values for each features
 - These vectors will be use to train a multi class model

CNN Architecture



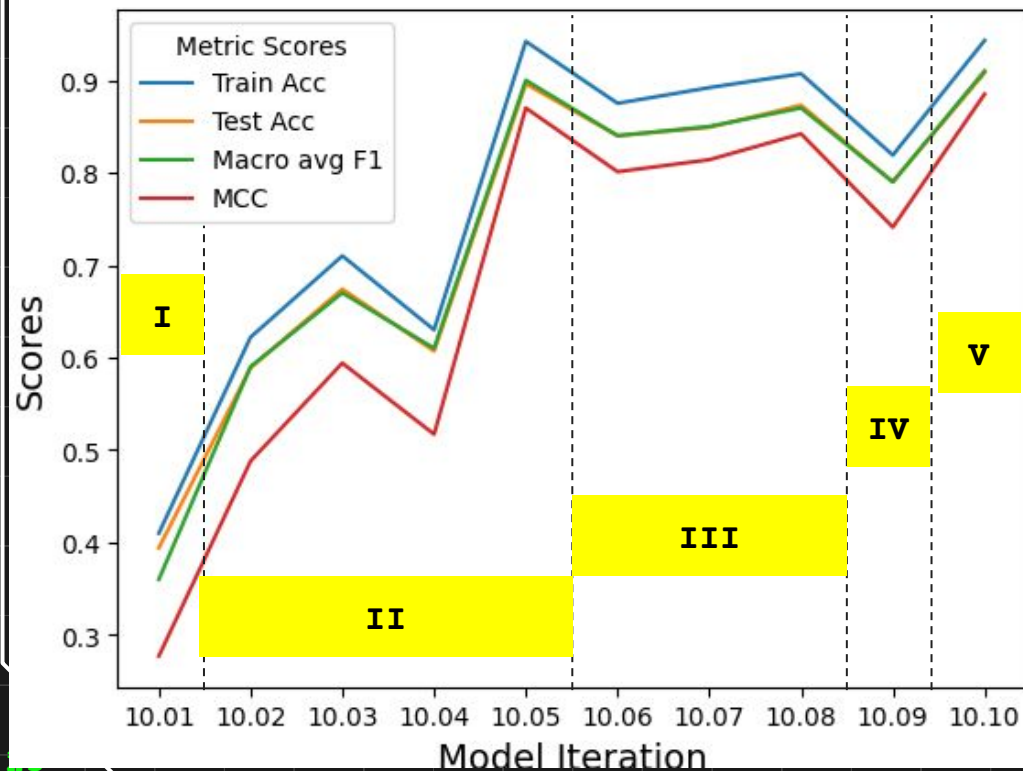
Layer info by colour

Conv(50) → Conv(50) ... Conv(75) → Conv(75) ... Conv(100) → Conv(100) ... Conv(150)

MaxPooling(2,2) → MaxPooling(2,2) → MaxPooling(1,2) → MaxPooling(1,2)

Dense(128) → Dense(64) → Dense(32) → Dense(5, Softmax)

Model Road Map



I) Baseline model

II) Adding more features

10.4 → 10.5 swap out features

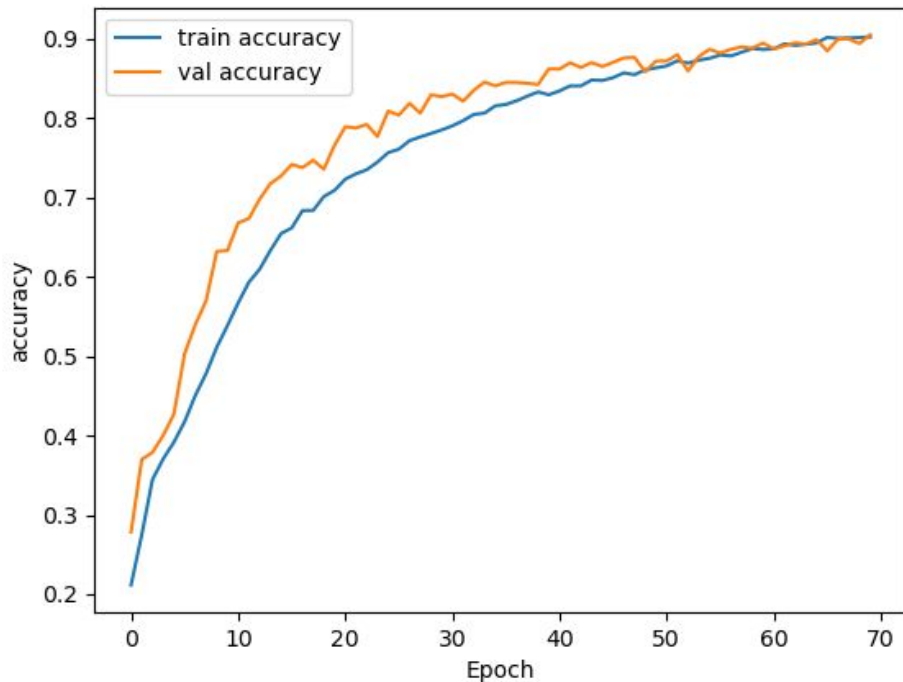
III) Adding more features

10.5 → 10.8 swap out features

IV) Adding last feature

V) Change Optimizer to Adamax from Adam

Best Model Evaluation



Train Acc: 0.943

Test Acc: 0.908

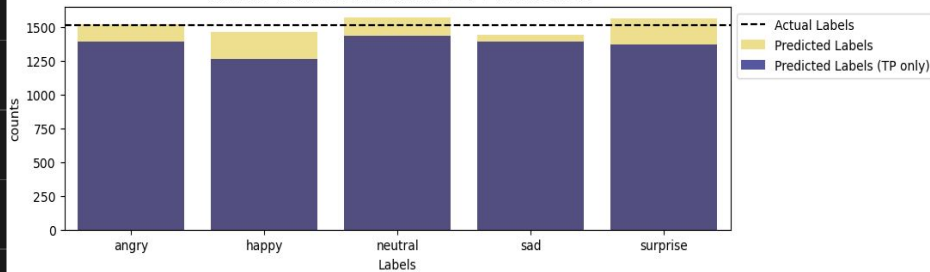
Macro average F1-Score: 0.910

MCC: 0.885

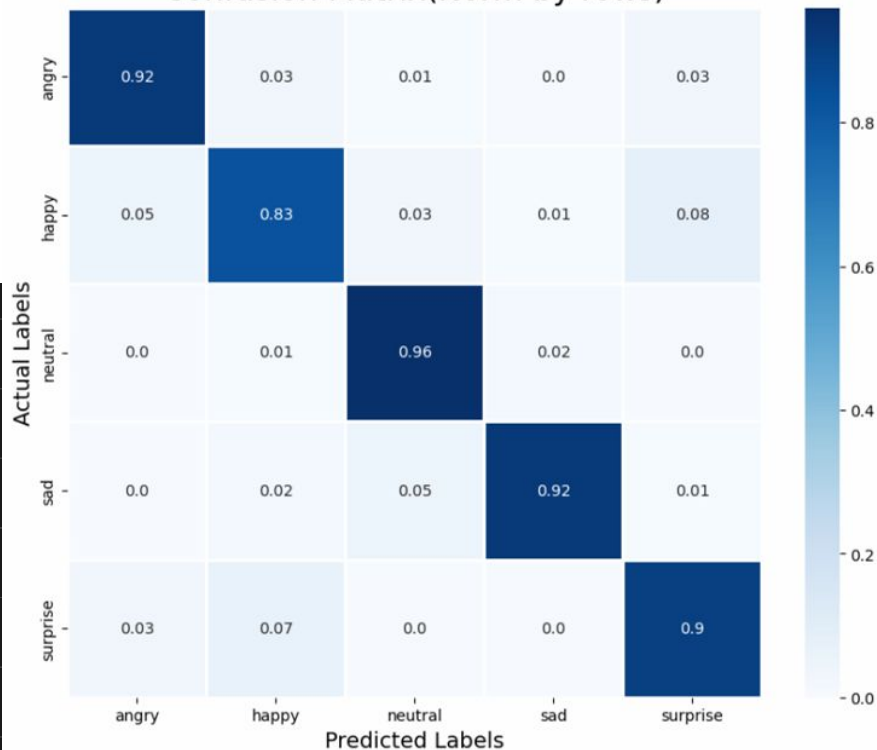
Model is not overfitted

Best Model Evaluation

Value counts Actual vs Predicted



Confusion Matrix(Norm by rows)



- Most misclassification are from happy & surprise
- Pleasant surprise can indistinguishable at times.
- Confusion Matrix Normalized by row (Recall)
- Recall for each class >80%

Application



Part 1: classifier

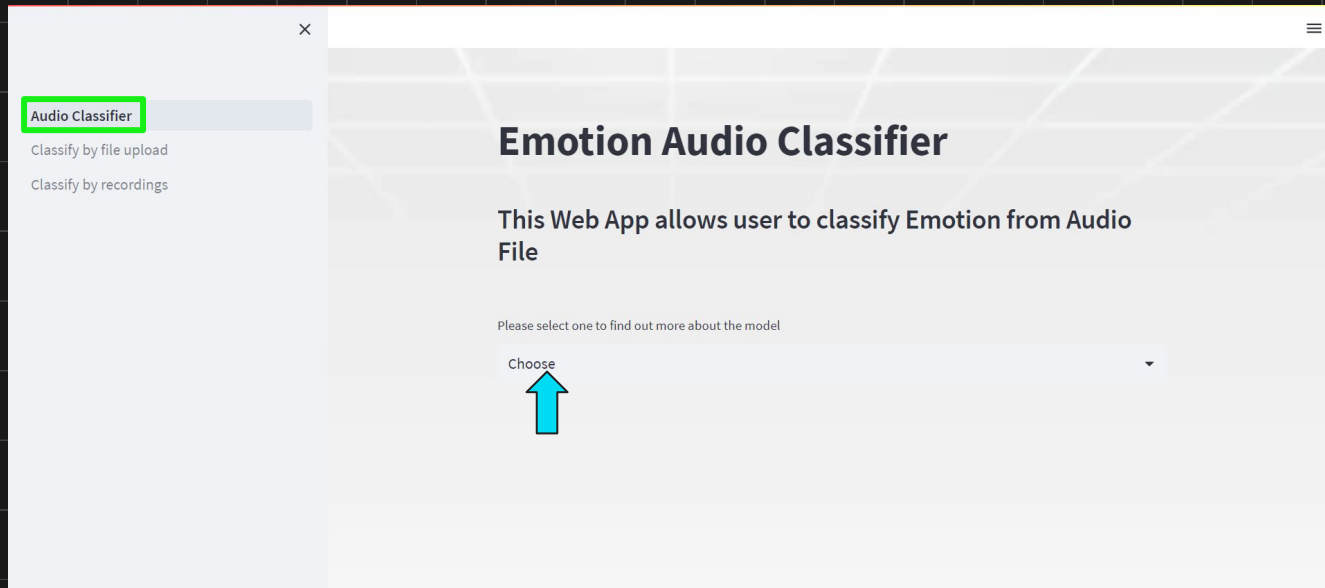
- Two ways to classify audio
 - ◆ Upload file
 - ◆ Record file
- File uploaded will be preprocessed for prediction
- Predicted Emotion will be displayed



Part2: Definition

- From Predicted class, get definition and examples.
- Video link for predicted emotion

Application (Main)



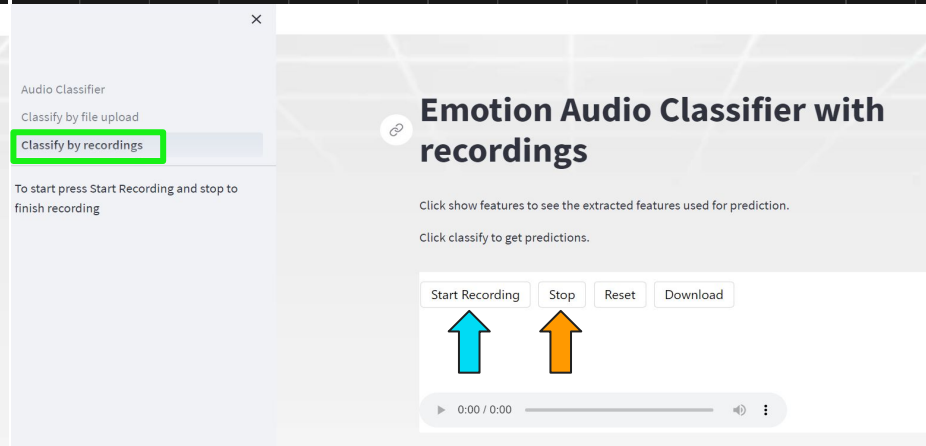
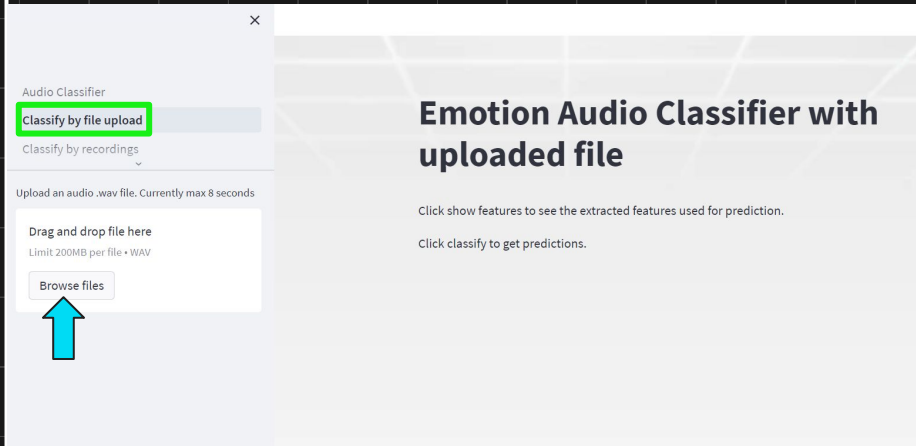
****Notice****

Web app not optimized for phone screen but is still usable

Application (Classifier)

File Upload route

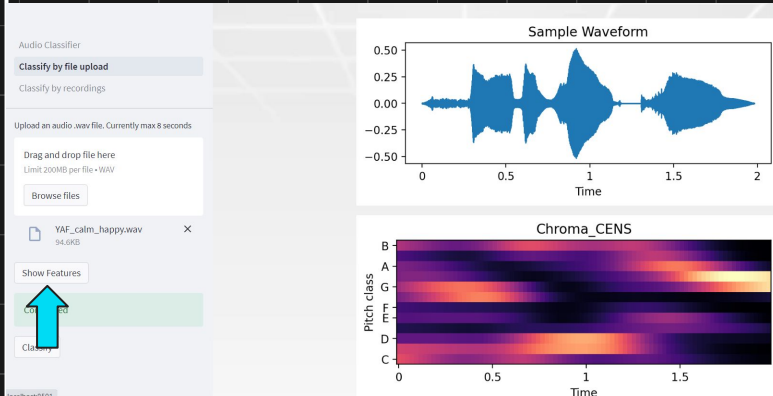
File Record route



Application (Classifier)

Show Features

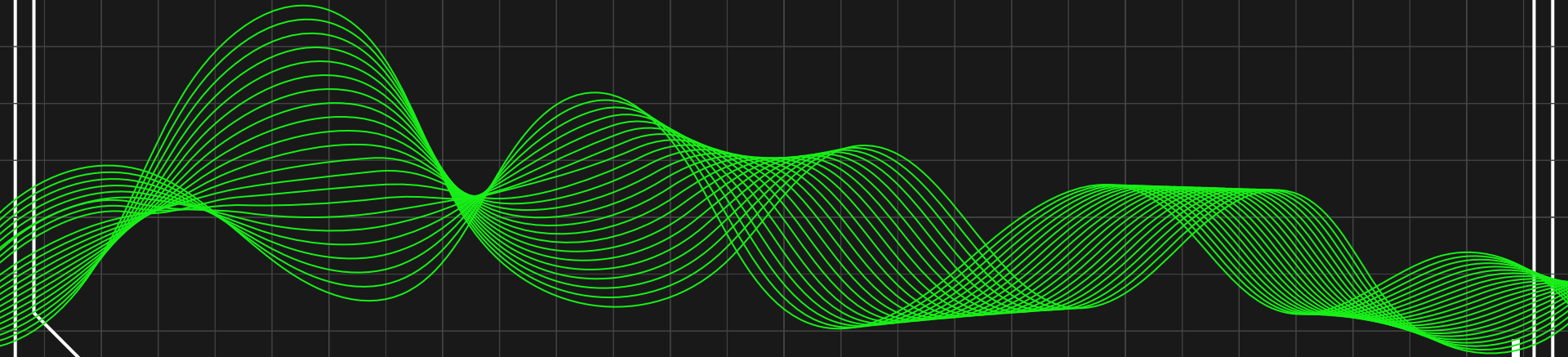
Classify



The 'Classify' interface shows the predicted emotion as 'ANGER'. Below this, there are four dropdown menus for 'Word Definition of Anger', 'Word Origin: Anger', 'Examples of Anger', and 'Video on Anger'. A green arrow points to the 'The predicted Emotion: ANGER' box, and four orange arrows point to the dropdown menus. A blue arrow points to the 'Classify' button, which is highlighted with a green bar and the text 'Classification completed'.



04// Conclusion & Recommendations



Conclusion

Best Model (CNN)

- Optimizer Adamax
- 6 audio features
- Step up conv2d layers
- Step down dense layers

Application Benefits

- Improved emotional awareness
- Better Social Intelligence
- Overall lifestyle change



Recommendations

Automation implementation

- Taking advantage of computer resources to carry out basic repetitive task

Deployment to Education sector

- Learning aid
- Promote social awareness

Integration to other services

- Networking systems where connecting and interaction with people with similar situations are possible.

Other Applications

- Customer services
- Healthcare
- Personal application

Limitations

5 Emotions

Current model only able to predict the 5 basic emotions
Anger, Happiness, Neutral,
Sadness & Surprise

Prediction are from whole audio

The audio are not segmented to analyse per word basis.

More to English speaking accent

Data used to train the model is mainly English speaking accent.

Short sentence

Model was trained with using short phrases instead of full sentences.

Thanks!

Do you have any questions?



<https://github.com/Ridzuan-M>



<https://www.linkedin.com/in/ridzuan-mokhtar>