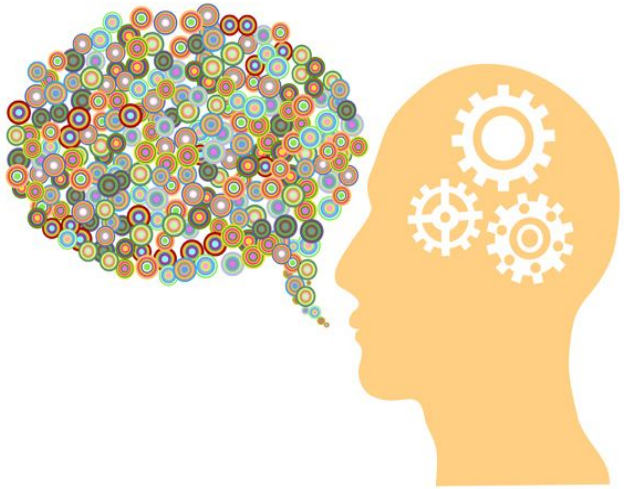
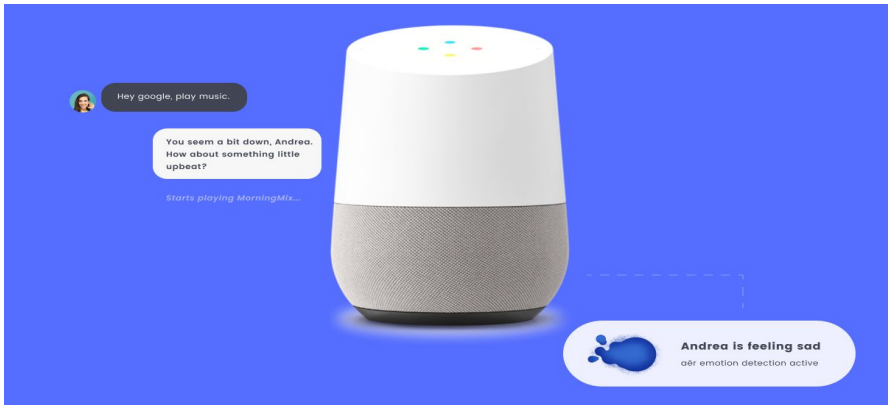
Foundations of Artificial Intelligence Project

Speech Emotion Recognition



INTRODUCTION

Emotions play a significant role in human mental life. It is a medium of expression of one's perspective or one's mental state to others. This is essential to our rational as well as intelligent decisions. It helps us to match and understand the feelings of others by conveying our feelings and giving feedback to others. Emotion awareness improves customer experience. This system helps machines understand the human emotions more effectively and get more emotion awareness.



Problem Statement

To understand the human emotions using two features – lexical and acoustic characteristics. Lexical characteristics are based on words and acoustic characteristics are based on the characteristics of the sound like the volume, pitch, quality etc.

Literature Survey

TABLE 1. Different Types of Features, Classifier and Dataset in Current Speech Emotional Recognition System

Ref	Types of classifier	Types of features	Recognition Rate	Type of Dataset	Methods
[10]	SVM	Prosodic and spectral features	44.4%	Berlin & LDC & FAU Aibo dataset	Ranking SVM
[18]	k-NN & linear discriminate	Fundamental frequency (F0), energy, duration, and the first and formant	40.7% for males & 36.4% for females	Private speech database from call center	domain-specific emotion recognition by k-NN and linear discriminate classifier
[16]	HMM, GMM, MLP and hierarchical model	Mean of the log-spectrum (MLS), MFCCs and prosodic features	HMM 68.57, Hierarchical model 71.75	Berlin dataset	Spectral characteristics of signals are used in order to group emotions based on acoustic rather than psychological considerations.
[11]	SVM	Zero crossing rate, root mean square energy, pitch, harmonics-to-noise ratio, and MFF	72.44% - 89.58%	AIBO and USC IEMOCAP dataset	Hierarchical computational structure to maps an input speech utterance into one of the multiple emotion classes
[20]	Bayesian Logistic Regression, SVM	large-margin feature	70.1% & 65.1% for two and five class	AIBO dataset Private Mandarin	Hierarchical structure for binary decision tree
[6]	G GMVAR	Mel-frequency cepstrum coefficient (MFCC),	76%	Berlin emotional speech database,	Gaussian mixture vector autoregressive (GMVAR) is a mixture of GMM with vector autoregressive for classification.
[22]	Binary classifier & QDC	Prosodic features such as energy contour and duration	75.8%	SEMAINE databases	Shape based method by using functional data analysis to obtain natural changeability of F0

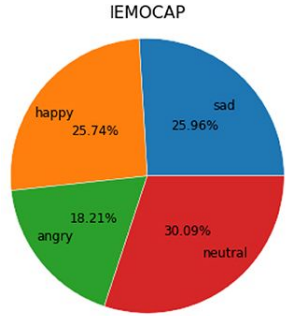
Datasets used

RAVDESS : Ryerson Audio - Visual Database of Emotional Speech and Song

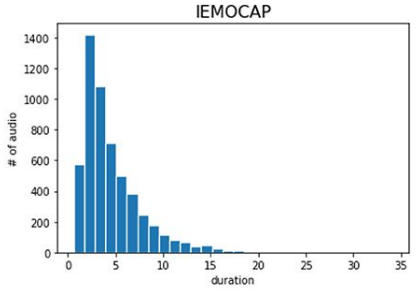
TESS: Toronto Emotional Speech Set

SAVEE: Surrey Audio-Visual Expressed Emotion

IEMOCAP: Interactive Emotional Dyadic Motion Capture Database

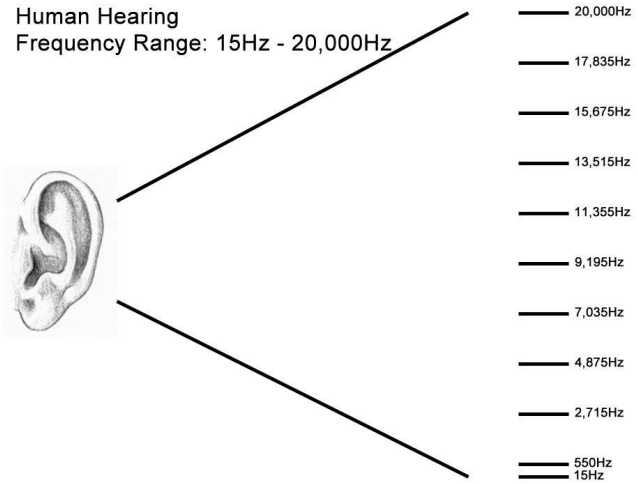


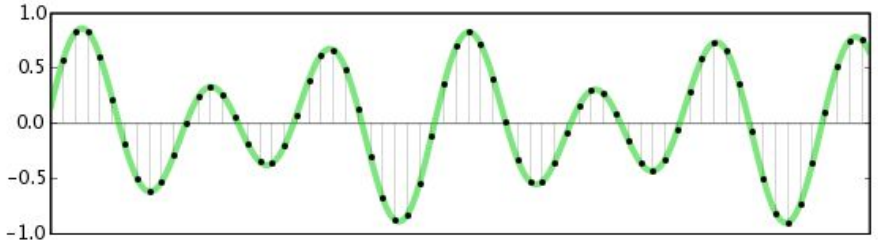
IEMOCAP label distribution



IEMOCAP audio length

Feature Extraction



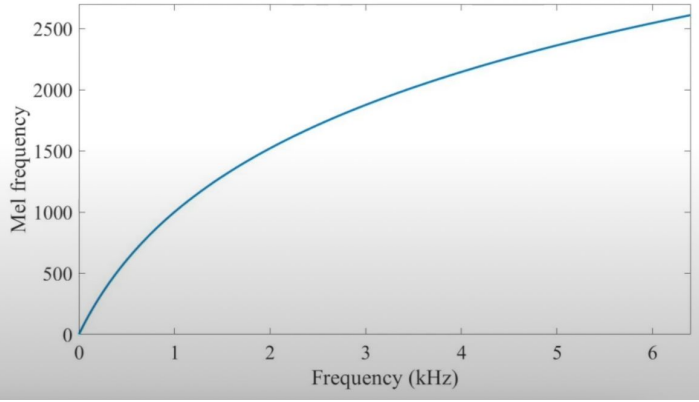


Mel Frequency

Humans perceive frequency logarithmically.

$$m = 2595 \cdot log(1 + \frac{f}{500})$$

$$f = 700(10^{m/2595} - 1)$$

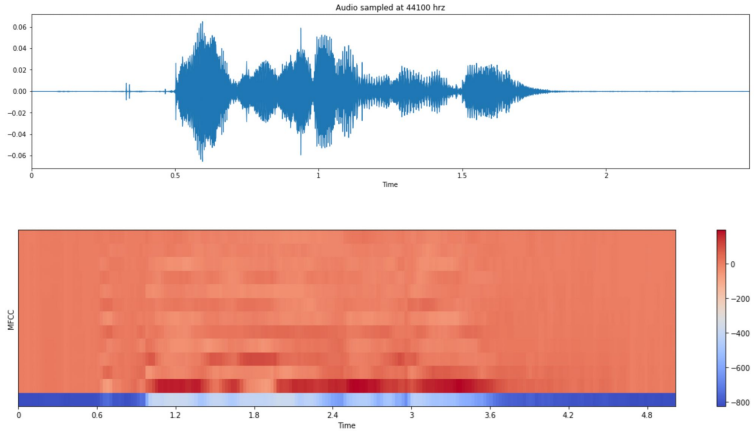


Launchpad

Feature Extraction

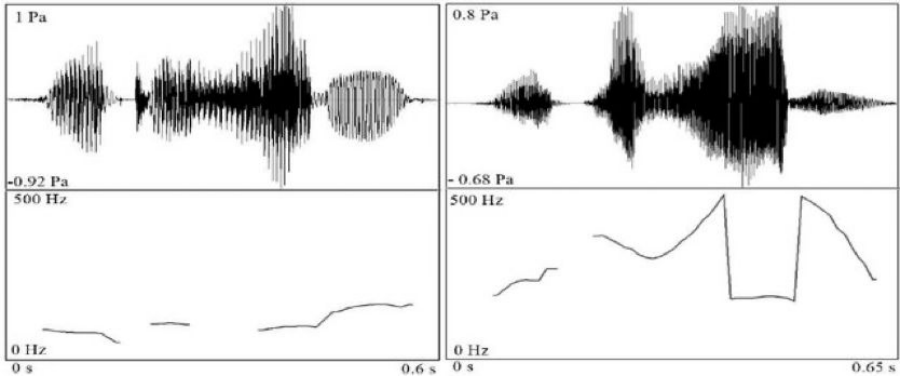
Mel Frequency Cepstral Coefficient (MFCC)

“The shape of the vocal manifests itself in the envelope of the short time power spectrum, and the job of the MFCC is to accurately represent this envelope.”

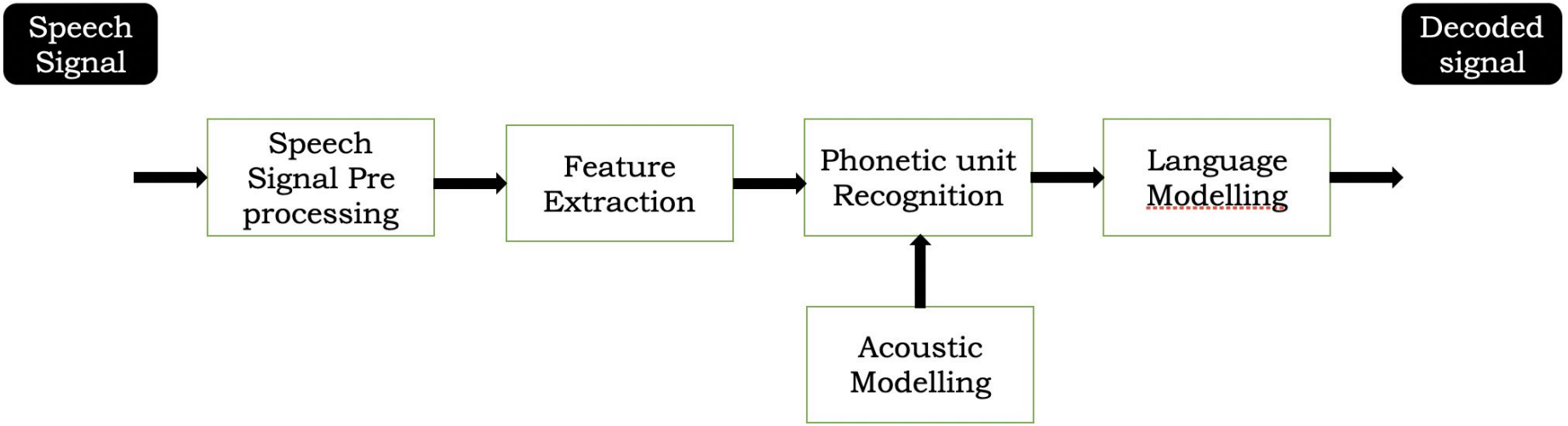


Feature Extraction

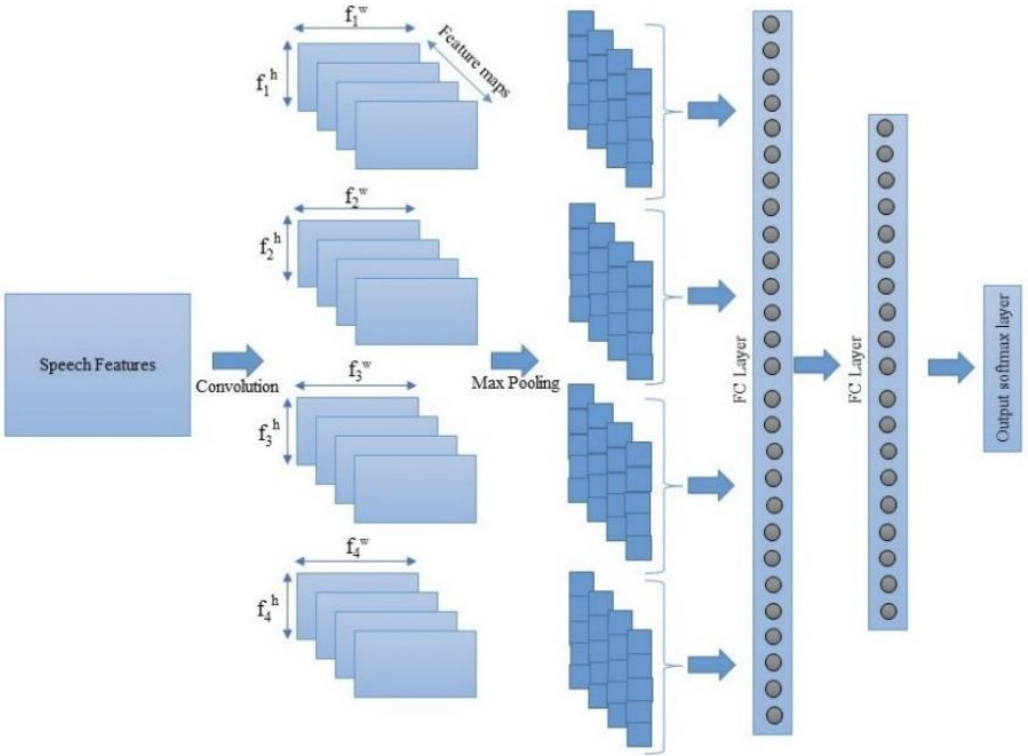
The top graphs represent the audio waves and the bottom graphs represent the pitch. The left hand side of the graph is normal speech and the right hand side is angry speech. The missing parts of the bottom graphs are parts of the speech signals which would not have foundation in human perception.



CNN Algorithm

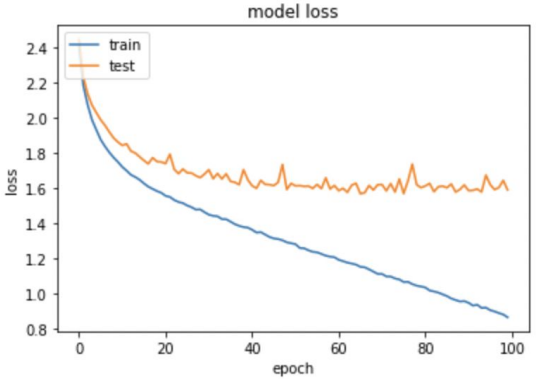


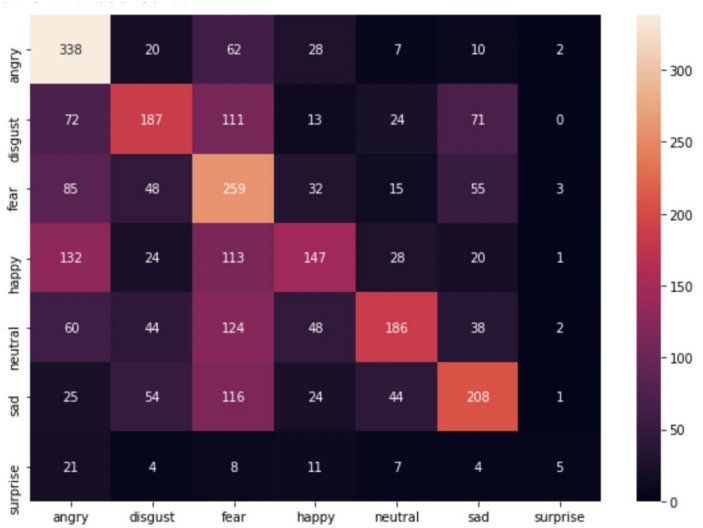
CNN Algorithm



Preliminary Results

The accuracy come to 76% using the CNN model.





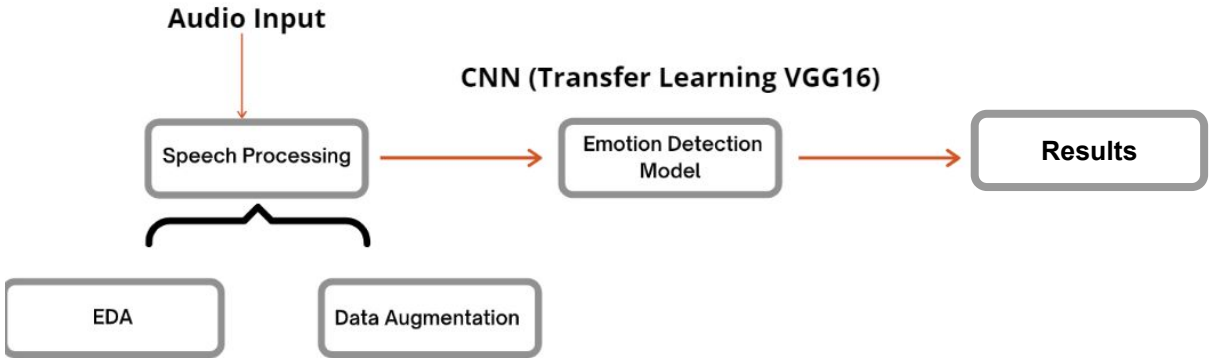
Preliminary Results

	actualvalues	predictedvalues
170	female_happy	female_happy
171	female_fear	female_fear
172	male_fear	female_happy
173	male_fear	female_angry
174	female_angry	female_fear
175	female_happy	female_neutral
176	female_happy	female_fear
177	female_fear	female_fear
178	female_neutral	female_neutral
179	female_angry	female_fear

predictedvalues	actualvalues
female_angry	562
female_disgust	335
female_fear	712
female_happy	270
female_neutral	247
female_sad	369
female_surprise	1
male_angry	171
male_disgust	46
male_fear	81
male_happy	33
male_neutral	64
male_sad	37
male_surprise	13

CNN Transfer learning using VGG16 model

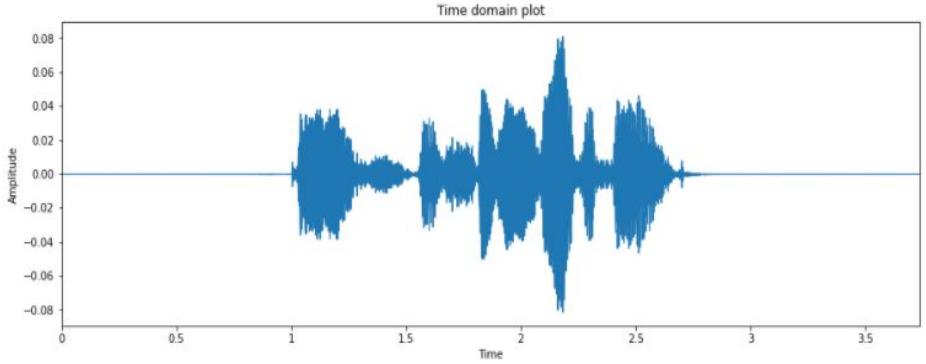
How the transfer learning model is implemented and the various models used?



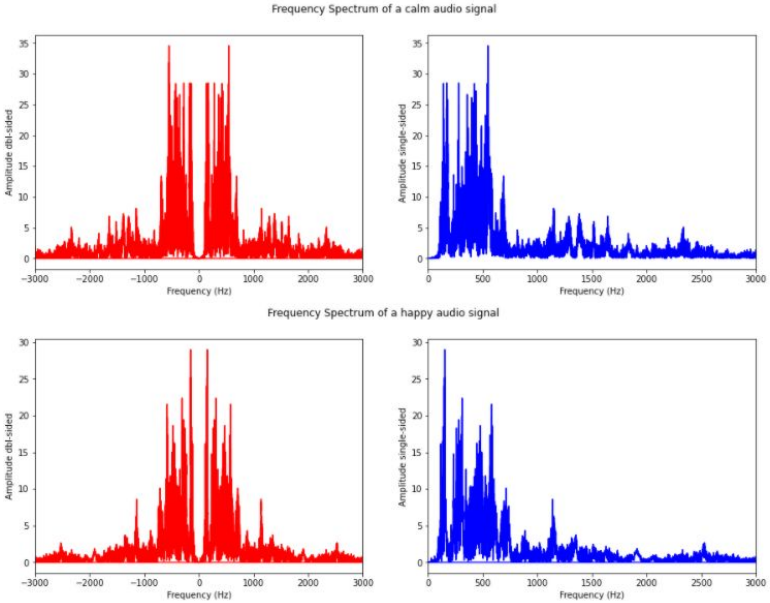
Speech Processing

Exploratory Data Analysis

Time domain



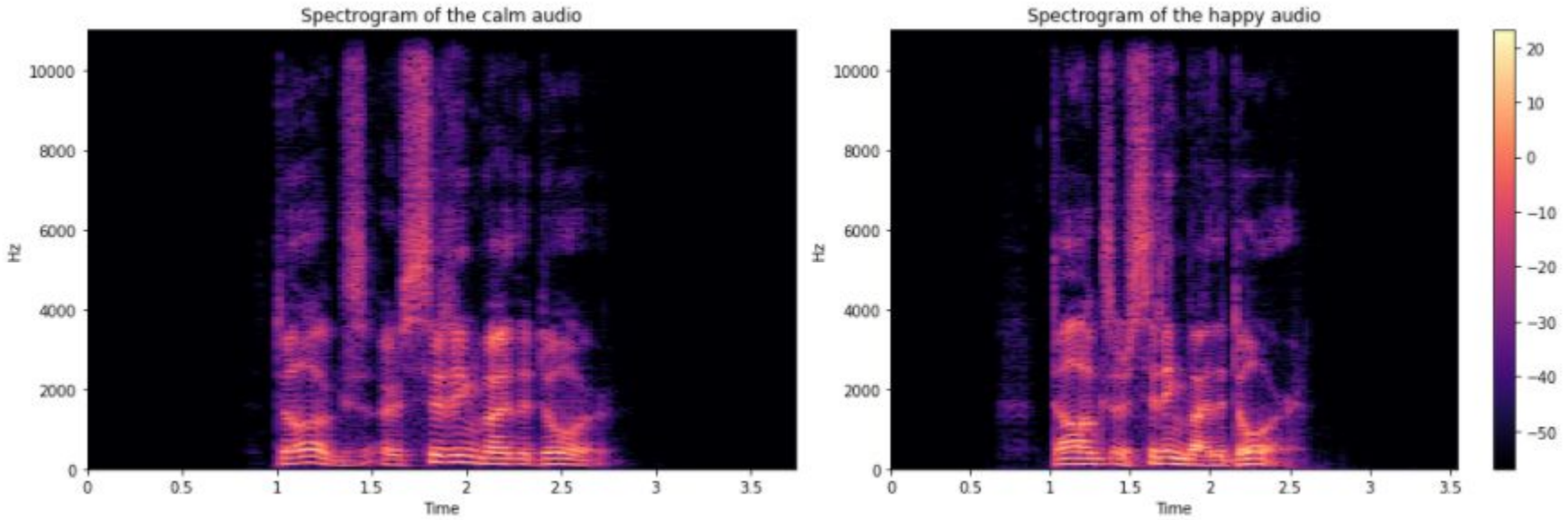
Frequency domain



Speech Processing

Exploratory Data Analysis

Time-frequency domain



Speech Processing

Data augmentation

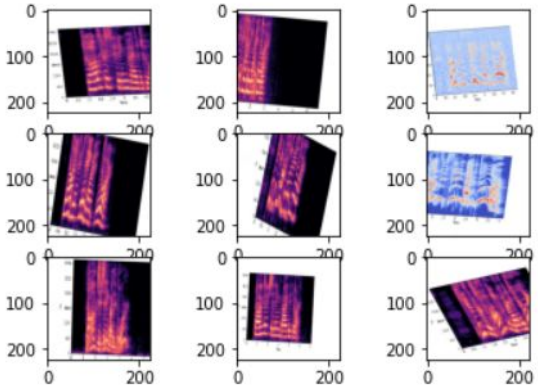
Signal Augmentation

Original Signal
Stretched Signal
White Noise added Signal
Compressed Signal

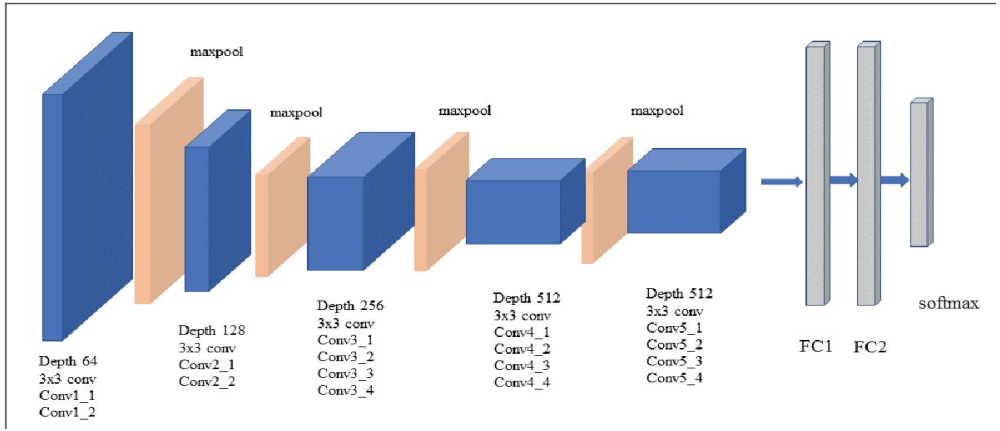
Spectrogram

Image Augmentation

Rotation
Zoom
Width Shift



Emotion Detector using Transfer Learning



Dataset used for Training: RAVDESS (Ryerson Audio - Visual Database of Emotional Speech and Song) **and SAVEE (**Surrey Audio-Visual Expressed Emotion)

VGG16 is a convolution neural net (CNN) architecture which was used to win ILSVR(Imagenet) competition in 2014. It is considered to be one of the excellent vision model architecture till date. Most unique thing about VGG16 is that instead of having a large number of hyper-parameter they focused on having convolution layers of 3x3 filter with a stride 1 and always used same padding and maxpool layer of 2x2 filter of stride 2.

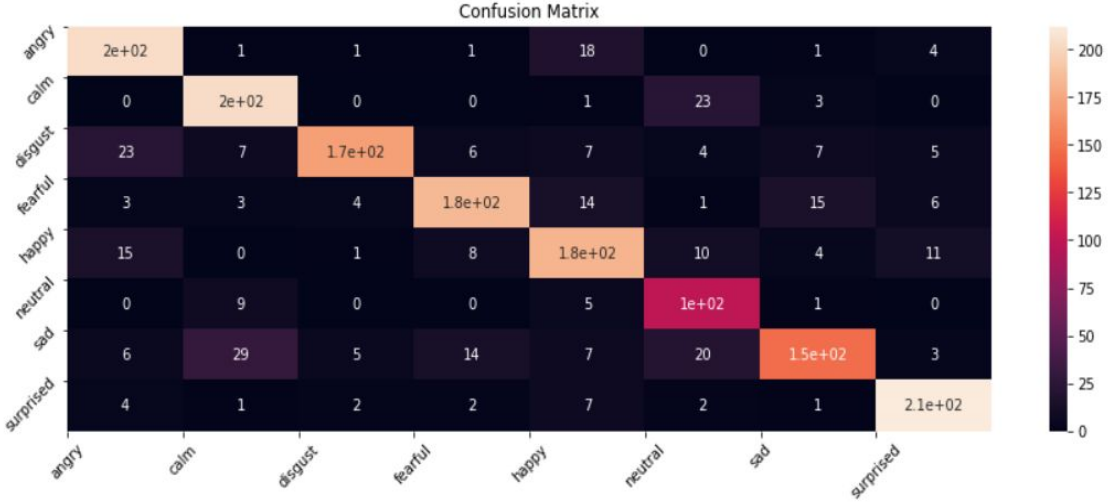
Model: "model"

Layer (type)	Output Shape	Param #
input_1 (InputLayer)	[(None, 224, 224, 3)]	0
block1_conv1 (Conv2D)	(None, 224, 224, 64)	1792
block1_conv2 (Conv2D)	(None, 224, 224, 64)	36928
block1_pool (MaxPooling2D)	(None, 112, 112, 64)	0
block2_conv1 (Conv2D)	(None, 112, 112, 128)	73856
block2_conv2 (Conv2D)	(None, 112, 112, 128)	147584
block2_pool (MaxPooling2D)	(None, 56, 56, 128)	0
block3_conv1 (Conv2D)	(None, 56, 56, 256)	295168
block3_conv2 (Conv2D)	(None, 56, 56, 256)	590080
block3_conv3 (Conv2D)	(None, 56, 56, 256)	590080
block3_pool (MaxPooling2D)	(None, 28, 28, 256)	0
block4_conv1 (Conv2D)	(None, 28, 28, 512)	1180160
block4_conv2 (Conv2D)	(None, 28, 28, 512)	2359808
block4_conv3 (Conv2D)	(None, 28, 28, 512)	2359808
block4_pool (MaxPooling2D)	(None, 14, 14, 512)	0
block5_conv1 (Conv2D)	(None, 14, 14, 512)	2359808
block5_conv2 (Conv2D)	(None, 14, 14, 512)	2359808
block5_conv3 (Conv2D)	(None, 14, 14, 512)	2359808
block5_pool (MaxPooling2D)	(None, 7, 7, 512)	0
flatten (Flatten)	(None, 25088)	0
dense (Dense)	(None, 512)	12845568
dropout (Dropout)	(None, 512)	0
dense_1 (Dense)	(None, 256)	131328
dense_2 (Dense)	(None, 8)	2056

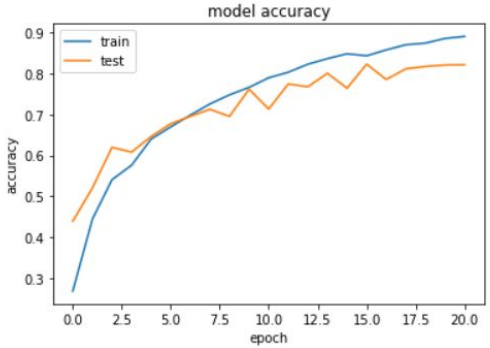
Total params: 27,693,640
Trainable params: 20,058,376
Non-trainable params: 7,635,264

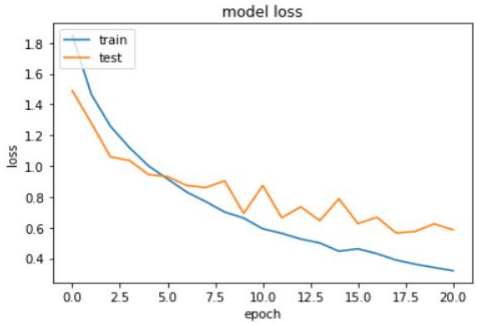
Our Results

Confusion matrix of VGG16 model detecting emotions



	precision	recall	f1-score	support
angry	0.80	0.89	0.84	230
calm	0.80	0.88	0.84	230
disgust	0.93	0.74	0.83	230
fearful	0.86	0.80	0.83	230
happy	0.76	0.79	0.77	231
neutral	0.62	0.87	0.73	115
sad	0.82	0.64	0.72	231
surprised	0.88	0.92	0.90	231
accuracy			0.81	1728
macro avg	0.81	0.82	0.81	1728
weighted avg	0.82	0.81	0.81	1728

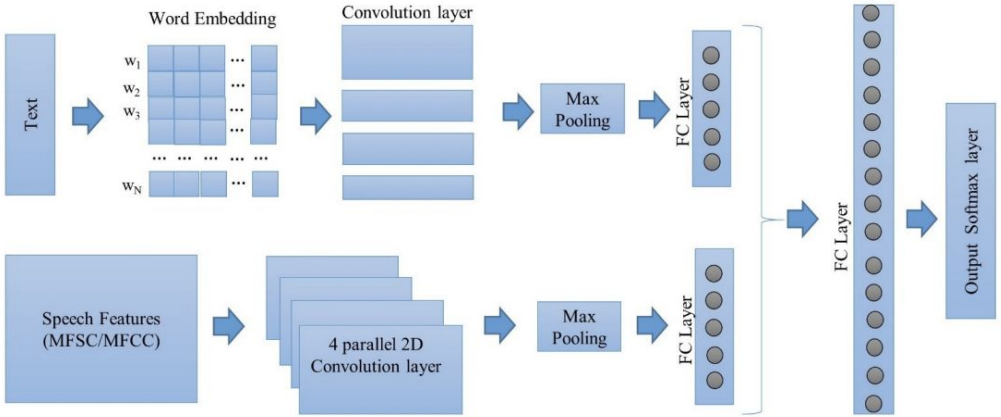





```
Detected Text : what the hell
Predicted Emotion: angry
Classwise Scores
Class  angry  has a score of  0.91
Class  calm  has a score of  0.0
Class  disgust  has a score of  0.0
Class  fearful  has a score of  0.09
Class  happy  has a score of  0.0
Class  neutral  has a score of  0.0
Class  sad  has a score of  0.0
Class  surprised  has a score of  0.0
```

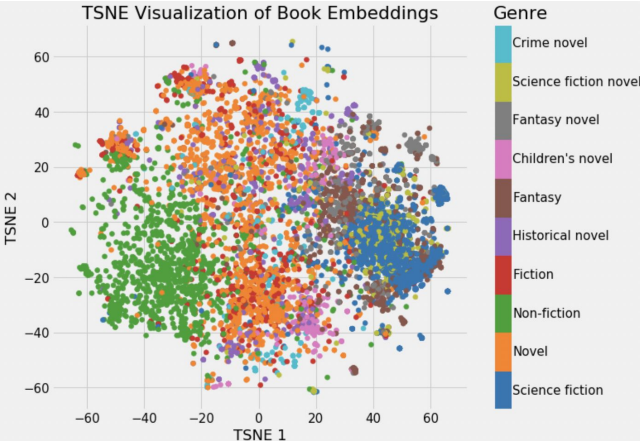
Multimodal Algorithm

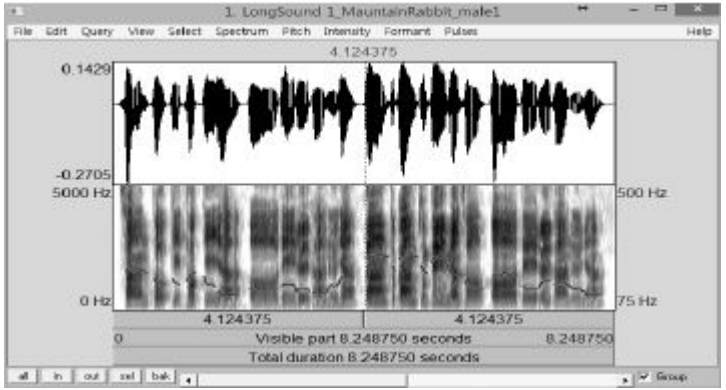
Along with the syntactic features, we combine the semantic features. MFCC and Spectrogram captures the low level features while text would also take the context of the sentence into consideration.



Feature Extraction

Word Embeddings, MFCC, Pitch Extraction





Further Work

- Completion of the multimodal speech recognition
- Implementation of attention based model to identify context from speech.
- Implementation of Therapy chatbot with interface.
- Application of the same in the area of Mental Health - to assess the tone of a person and accordingly provide audio responses catered to the mental health issue

References

- Medium : Speech Emotion Recognition With Convolutional Neural Network by Diego Rios
- ‘Deep Learning based Emotion Recognition System Using Speech Features and Transcriptions’ by Suraj Triapthi, Abhay Kumar, Abhiram Ramesh, Chirag Singh and Pramod Yenigalla
- Medium: Extract Features, Visualize Filters and Feature Maps in VGG16 and VGG19 CNN Models
- Attention Based Fully Convolutional Network for Speech Emotion Recognition by Yuanyuan Zhang et al.