Project 4-2

Subject:

Emotion Recognition on Audio

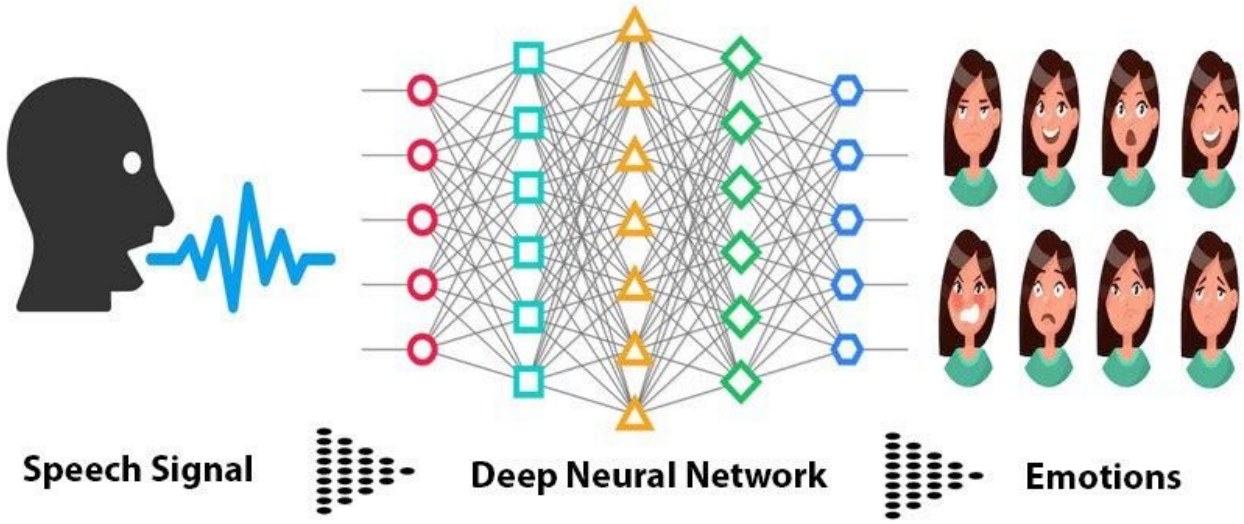
Authors:

Zahra MohammadBeigi

Student No. : 97222079

Ghazal Rafiei

Student No. : 97222044



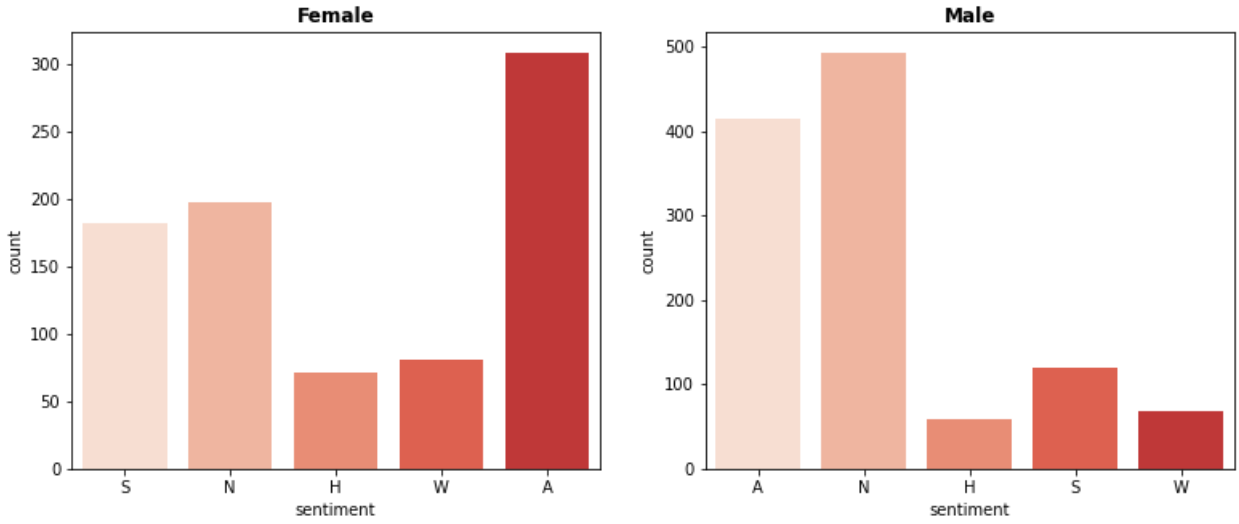
Introduction

Emotion recognition from natural sounds is a very challenging problem. The audio sub-challenge represents an initial step towards building an efficient audio based emotion recognition system that can detect emotions for real life applications (i.e. human-machine interaction and/or communication).

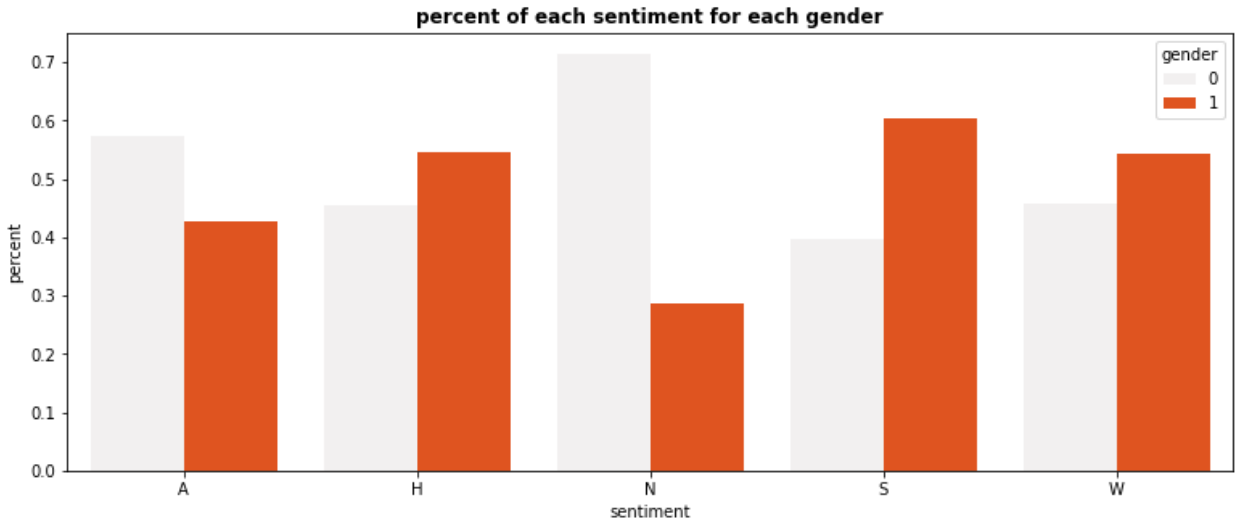
The data provided of audio cannot be understood by the models directly.to convert them into an understandable format feature extraction is used.It is a process that explains most of the data but in an understandable way. In this project we used ***Mel-frequency cepstral coefficients (MFCC)*** and ***Mel-Spectrogram***.

Get Some Information from Dataset and Visualizations

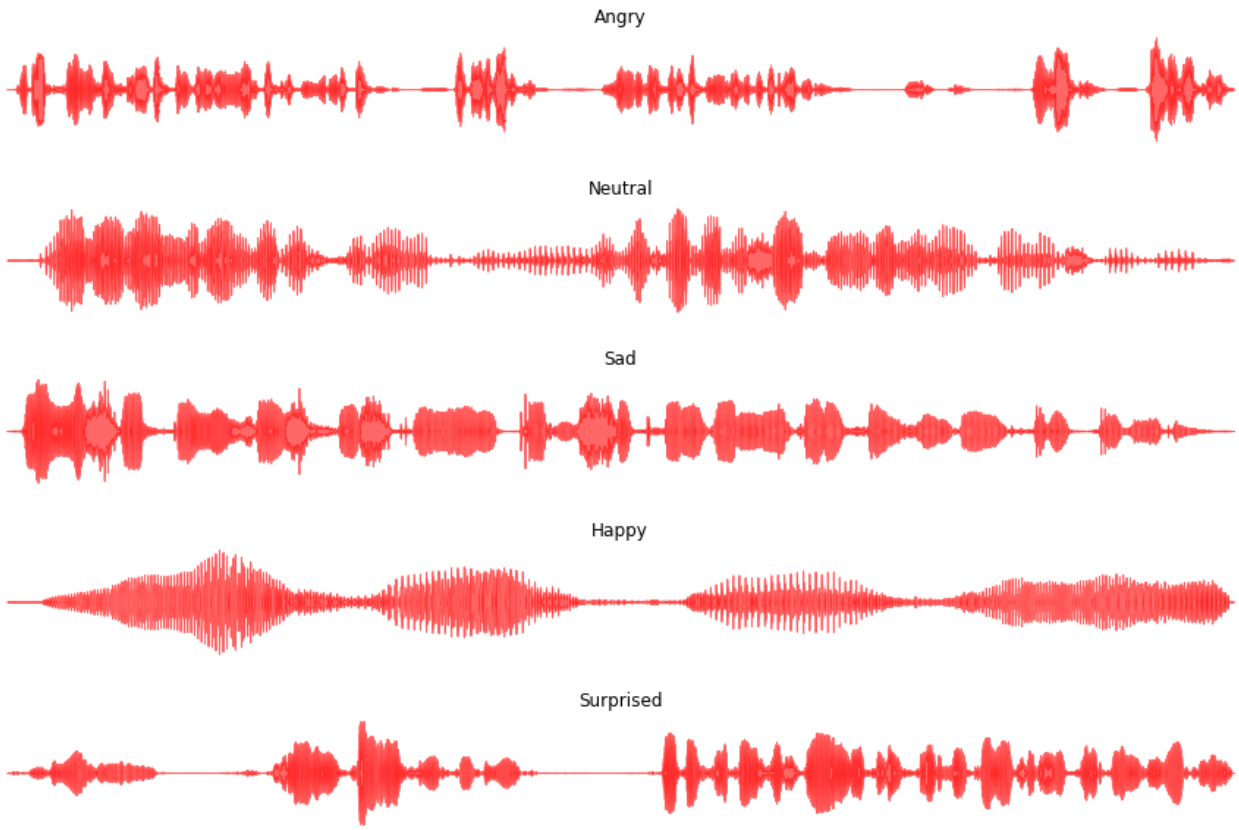
In our train dataset we have ***1994*** strings of length 10. The first four characters represent the id of the audio and the fifth one represents the gender and finally the sixth represents the type of emotion. We separate these information in id, gender, and sentiment lists respectively. For example the string ***1251MN.wav*** indicates that the audio with the id *1251* is the sound of a *Neutral man*. Our audio file consists of 1155 men sounds and 839 women sounds and this is the distribution over those genders.



Here we got the distribution of each sentiment over each gender(White is for men and orange is for women). For example there are more angry men than women and more sad women than men.



Here you can see the pure visualization of waves of 5 different sentiments:



Preprocessing

Feature extracting:

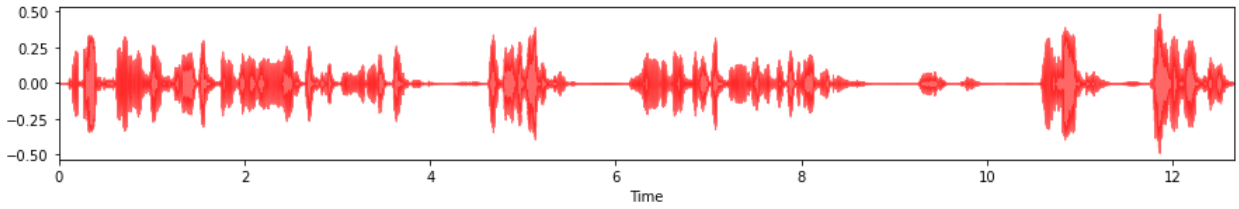
There are many features that can be extracted from a single sound, but there are some that are used in emotion recognition. For example there are others which are used in speech recognition.

Mostly, in sentiment analysis, MFCC and Mel-spectrogram which you can see the impact below.

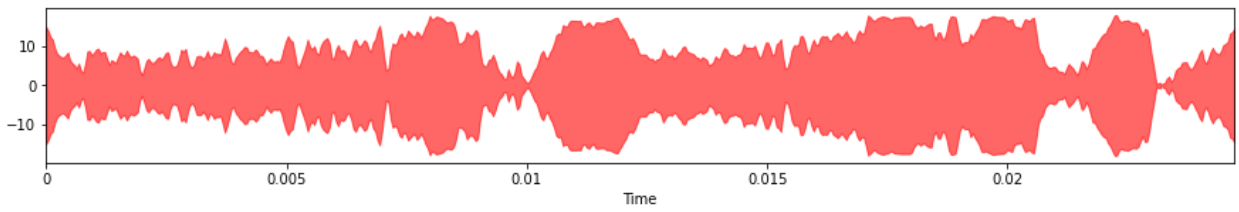
Before MFCC:

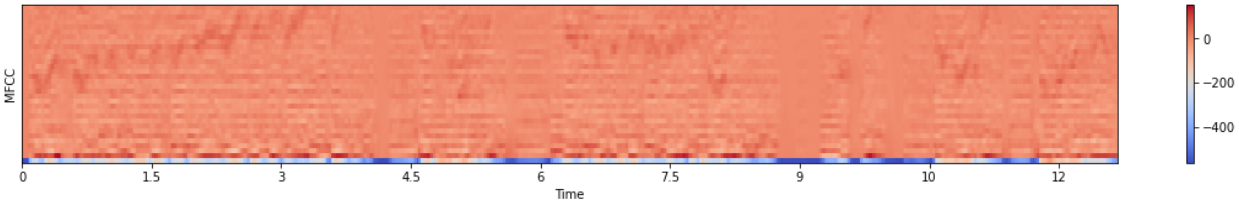
Let's see what mfcc do to the pure sound:

It is good to mention there is a parameter called n_mfcc in feature extraction which we changed to 16, ,32, 64 and it didn't make a difference so we ended with using 32.



After MFCC:





This diagram describes the overall shape of a spectral envelope. It actually models the characteristics of the human voice.

After Feature extracting we normalized the data and then there was a problem for feeding it to the network:

Sounds did not have the same length so we set a fixed size which is the average of all sounds which is 180 and enforced the sounds to exactly have that length by adding zero or removing some numbers from the end.

Training

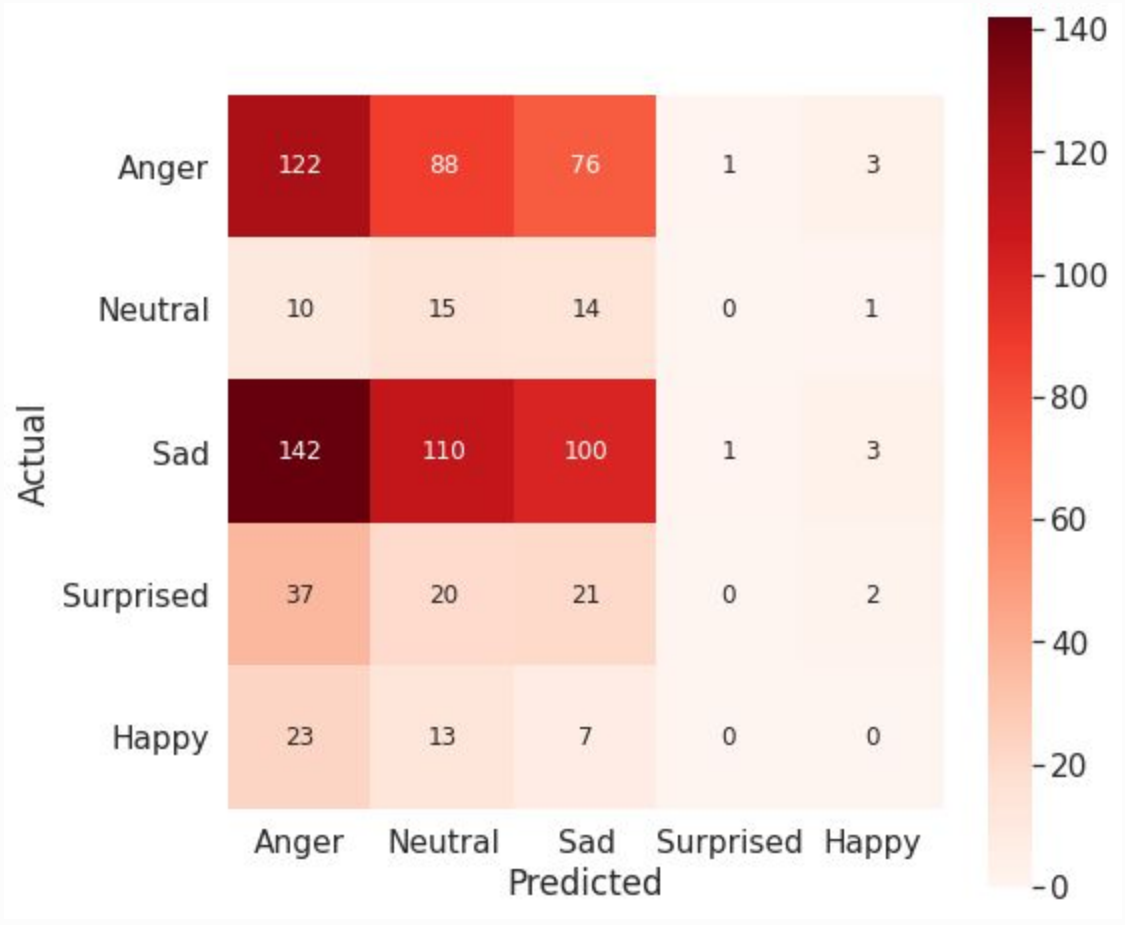
It seemed separating men and women could not help us because after all, we will have test cases from both classes and it's not appropriate to omit one class at all. Although, we tested separating and you can see the result. After that we kept going on the whole dataset.

MFCC-only on men:

```
Epoch 1/1000
18/18 - 3s - loss: 1.9549 - accuracy: 0.2023
Epoch 2/1000
18/18 - 0s - loss: 1.7693 - accuracy: 0.2428
Epoch 3/1000
18/18 - 0s - loss: 1.6415 - accuracy: 0.2948
Epoch 4/1000
18/18 - 0s - loss: 1.5943 - accuracy: 0.3266
Epoch 5/1000
18/18 - 1s - loss: 1.4715 - accuracy: 0.3873 - val_loss: 1.4989 - val_accuracy: 0.4227
Epoch 6/1000
18/18 - 0s - loss: 1.5054 - accuracy: 0.4075
Epoch 7/1000
18/18 - 0s - loss: 1.4421 - accuracy: 0.4306
Epoch 8/1000
18/18 - 0s - loss: 1.3477 - accuracy: 0.4711
Epoch 9/1000
18/18 - 0s - loss: 1.4044 - accuracy: 0.4538
Epoch 10/1000
18/18 - 0s - loss: 1.4246 - accuracy: 0.4075 - val_loss: 1.4643 - val_accuracy: 0.4499
Epoch 11/1000
18/18 - 0s - loss: 1.3549 - accuracy: 0.4566
Epoch 12/1000
18/18 - 0s - loss: 1.3374 - accuracy: 0.4798
Epoch 13/1000
18/18 - 0s - loss: 1.2840 - accuracy: 0.4798
Epoch 14/1000
18/18 - 0s - loss: 1.2999 - accuracy: 0.5058
Epoch 15/1000
18/18 - 0s - loss: 1.2403 - accuracy: 0.5029 - val_loss: 1.3310 - val_accuracy: 0.4363
Epoch 16/1000
18/18 - 0s - loss: 1.2438 - accuracy: 0.5000
Epoch 17/1000
18/18 - 0s - loss: 1.2337 - accuracy: 0.5231
Epoch 18/1000
18/18 - 0s - loss: 1.2266 - accuracy: 0.5289
Epoch 19/1000
18/18 - 0s - loss: 1.2391 - accuracy: 0.4942
Epoch 20/1000
18/18 - 0s - loss: 1.2353 - accuracy: 0.5000 - val_loss: 1.3151 - val_accuracy: 0.4722
```

```
Epoch 66/1000
18/18 - 0s - loss: 0.9687 - accuracy: 0.6358
Epoch 67/1000
18/18 - 0s - loss: 0.9612 - accuracy: 0.6243
Epoch 68/1000
18/18 - 0s - loss: 0.9999 - accuracy: 0.6156
Epoch 69/1000
18/18 - 0s - loss: 1.0142 - accuracy: 0.5983
Epoch 70/1000
18/18 - 0s - loss: 1.0018 - accuracy: 0.5925 - val_loss: 1.3582 - val_accuracy: 0.4314
Epoch 71/1000
18/18 - 0s - loss: 1.0072 - accuracy: 0.6098
Epoch 72/1000
18/18 - 0s - loss: 0.9886 - accuracy: 0.6040
Epoch 73/1000
18/18 - 0s - loss: 1.0645 - accuracy: 0.5838
Epoch 74/1000
18/18 - 0s - loss: 1.0090 - accuracy: 0.6098
Epoch 75/1000
18/18 - 0s - loss: 0.9743 - accuracy: 0.6069 - val_loss: 1.5797 - val_accuracy: 0.3832
Epoch 76/1000
18/18 - 0s - loss: 0.9721 - accuracy: 0.6185
Epoch 77/1000
18/18 - 0s - loss: 0.9589 - accuracy: 0.6243
Epoch 78/1000
18/18 - 0s - loss: 0.9806 - accuracy: 0.6214
Epoch 79/1000
18/18 - 0s - loss: 0.9616 - accuracy: 0.6243
Epoch 80/1000
18/18 - 0s - loss: 0.9948 - accuracy: 0.5694 - val_loss: 1.4819 - val_accuracy: 0.4363
Epoch 81/1000
18/18 - 0s - loss: 1.0715 - accuracy: 0.5636
Epoch 82/1000
18/18 - 0s - loss: 1.0310 - accuracy: 0.6040
Epoch 83/1000
18/18 - 0s - loss: 1.0498 - accuracy: 0.5925
Epoch 84/1000
18/18 - 0s - loss: 1.0476 - accuracy: 0.5607
Epoch 85/1000
18/18 - 0s - loss: 1.0054 - accuracy: 0.6098 - val_loss: 1.3121 - val_accuracy: 0.4759
Epoch 86/1000
18/18 - 0s - loss: 0.9842 - accuracy: 0.5983
Epoch 00086: early stopping
```

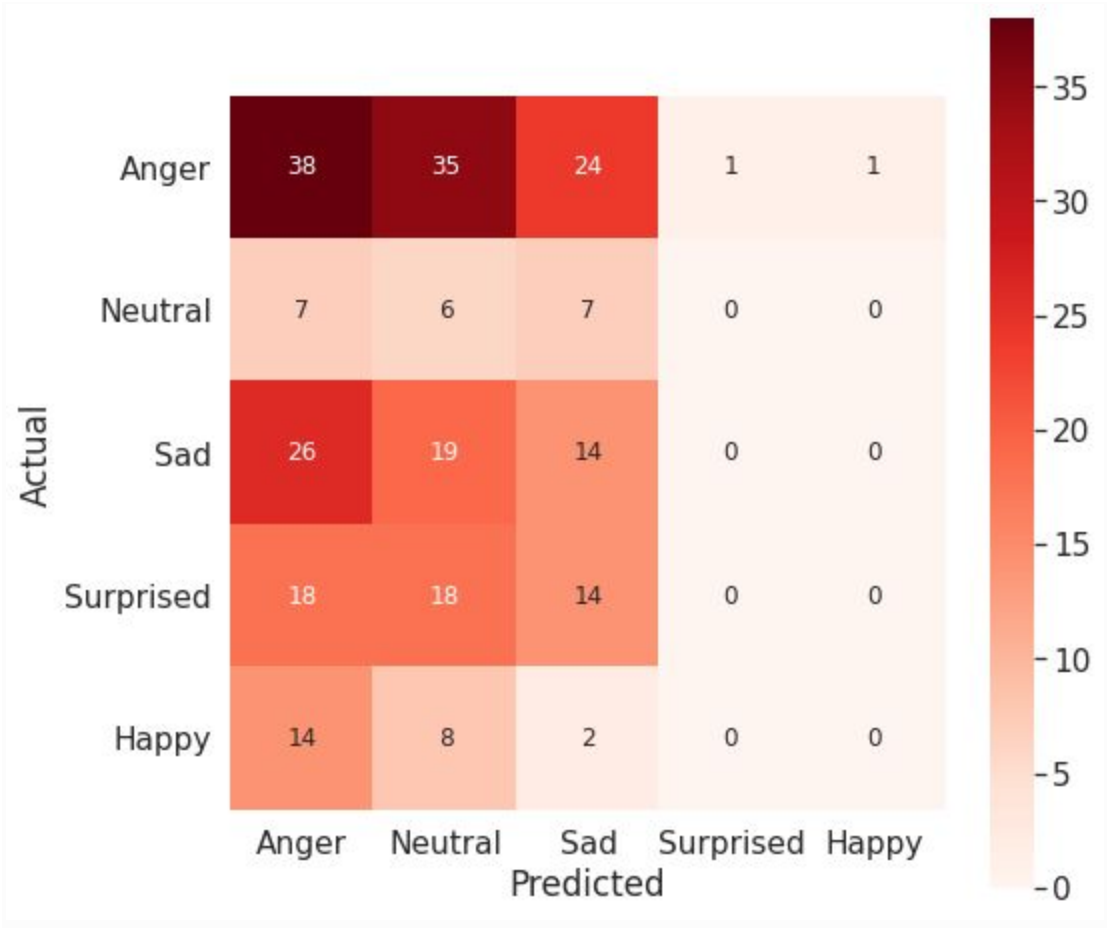
Confusion Matrix of Predicting and Actual Labels:



MFCC-only on women:

```
Epoch 1/1000
30/30 - 0s - loss: 1.7986 - accuracy: 0.3118
Epoch 2/1000
30/30 - 0s - loss: 1.5902 - accuracy: 0.3339
Epoch 3/1000
30/30 - 0s - loss: 1.5066 - accuracy: 0.3612
Epoch 4/1000
30/30 - 0s - loss: 1.4469 - accuracy: 0.3884
Epoch 5/1000
30/30 - 0s - loss: 1.4505 - accuracy: 0.3714 - val_loss: 1.4346 - val_accuracy: 0.3571
Epoch 6/1000
30/30 - 0s - loss: 1.4306 - accuracy: 0.3799
Epoch 7/1000
30/30 - 0s - loss: 1.3890 - accuracy: 0.4276
Epoch 8/1000
30/30 - 0s - loss: 1.4037 - accuracy: 0.4037
Epoch 9/1000
30/30 - 0s - loss: 1.3760 - accuracy: 0.4174
Epoch 10/1000
30/30 - 0s - loss: 1.3608 - accuracy: 0.4055 - val_loss: 1.4127 - val_accuracy: 0.3571
Epoch 11/1000
30/30 - 0s - loss: 1.3339 - accuracy: 0.4566
Epoch 12/1000
30/30 - 0s - loss: 1.3325 - accuracy: 0.4378
Epoch 13/1000
30/30 - 0s - loss: 1.3401 - accuracy: 0.4395
Epoch 14/1000
30/30 - 0s - loss: 1.3361 - accuracy: 0.4566
Epoch 15/1000
30/30 - 0s - loss: 1.3185 - accuracy: 0.4446 - val_loss: 1.4002 - val_accuracy: 0.4167
Epoch 16/1000
30/30 - 0s - loss: 1.3271 - accuracy: 0.4532
Epoch 17/1000
30/30 - 0s - loss: 1.2968 - accuracy: 0.4617
Epoch 18/1000
30/30 - 0s - loss: 1.2916 - accuracy: 0.4634
Epoch 19/1000
30/30 - 0s - loss: 1.2941 - accuracy: 0.4651
```

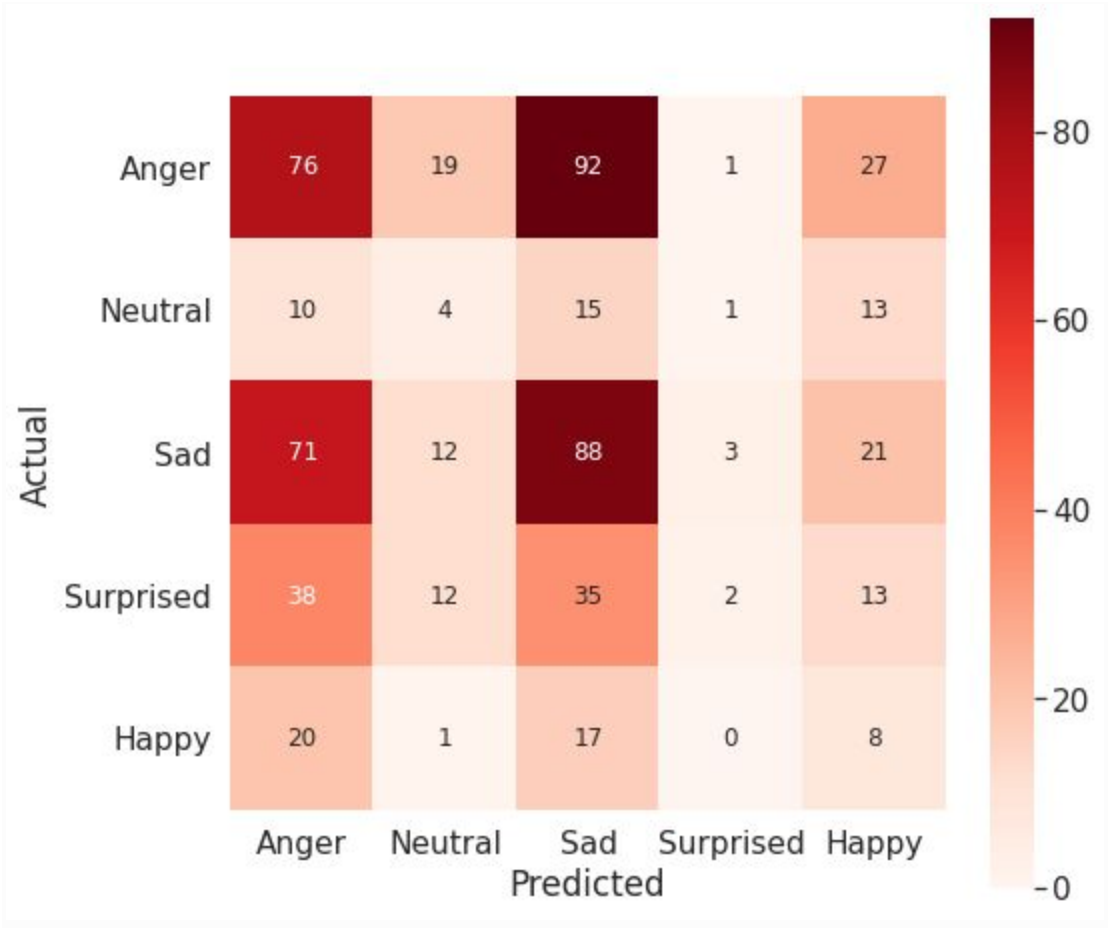
```
Epoch 138/1000
30/30 - 0s - loss: 0.4986 - accuracy: 0.8211
Epoch 139/1000
30/30 - 0s - loss: 0.5347 - accuracy: 0.8092
Epoch 140/1000
30/30 - 0s - loss: 0.5264 - accuracy: 0.7922 - val_loss: 2.4002 - val_accuracy: 0.3452
Epoch 141/1000
30/30 - 0s - loss: 0.4773 - accuracy: 0.8228
Epoch 142/1000
30/30 - 0s - loss: 0.6193 - accuracy: 0.7734
Epoch 143/1000
30/30 - 0s - loss: 0.4766 - accuracy: 0.8365
Epoch 144/1000
30/30 - 0s - loss: 0.4679 - accuracy: 0.8416
Epoch 145/1000
30/30 - 0s - loss: 0.4588 - accuracy: 0.8433 - val_loss: 2.3625 - val_accuracy: 0.3770
Epoch 146/1000
30/30 - 0s - loss: 0.4151 - accuracy: 0.8654
Epoch 147/1000
30/30 - 0s - loss: 0.4345 - accuracy: 0.8467
Epoch 148/1000
30/30 - 0s - loss: 0.4975 - accuracy: 0.8245
Epoch 149/1000
30/30 - 0s - loss: 0.5062 - accuracy: 0.8177
Epoch 150/1000
30/30 - 0s - loss: 0.4747 - accuracy: 0.8416 - val_loss: 2.4837 - val_accuracy: 0.3373
Epoch 151/1000
30/30 - 0s - loss: 0.4392 - accuracy: 0.8501
Epoch 152/1000
30/30 - 0s - loss: 0.4276 - accuracy: 0.8671
Epoch 153/1000
30/30 - 0s - loss: 0.4842 - accuracy: 0.8365
Epoch 154/1000
30/30 - 0s - loss: 0.4452 - accuracy: 0.8433
Epoch 00154: early stopping
```


Anger
Neutral
Sad
Surprised
Happy
Actual
38
35
24
1
1
7
6
7
0
0
26
19
14
0
0
18
18
14
0
0
14
8
2
0
0
Anger
Neutral
Sad
Surprised
Happy
Predicted
35
30
25
20
15
10
5
0

MFCC- on the whole:

```
Epoch 1/1000
70/70 - 0s - loss: 1.6884 - accuracy: 0.4208
Epoch 2/1000
70/70 - 0s - loss: 1.3564 - accuracy: 0.4710
Epoch 3/1000
70/70 - 0s - loss: 1.2980 - accuracy: 0.4932
Epoch 4/1000
70/70 - 0s - loss: 1.2913 - accuracy: 0.4946
Epoch 5/1000
70/70 - 0s - loss: 1.2923 - accuracy: 0.4889 - val_loss: 1.3699 - val_accuracy: 0.4190
Epoch 6/1000
70/70 - 0s - loss: 1.2574 - accuracy: 0.4961
Epoch 7/1000
70/70 - 0s - loss: 1.2641 - accuracy: 0.4817
Epoch 8/1000
70/70 - 0s - loss: 1.2457 - accuracy: 0.4968
Epoch 9/1000
70/70 - 0s - loss: 1.2066 - accuracy: 0.5147
Epoch 10/1000
70/70 - 0s - loss: 1.2232 - accuracy: 0.5204 - val_loss: 1.3922 - val_accuracy: 0.4140
Epoch 11/1000
70/70 - 0s - loss: 1.2186 - accuracy: 0.5111
Epoch 12/1000
70/70 - 0s - loss: 1.1973 - accuracy: 0.5276
Epoch 13/1000
70/70 - 0s - loss: 1.1831 - accuracy: 0.5412
Epoch 14/1000
70/70 - 0s - loss: 1.1859 - accuracy: 0.5491
Epoch 15/1000
70/70 - 0s - loss: 1.1784 - accuracy: 0.5391 - val_loss: 1.3784 - val_accuracy: 0.4691
Epoch 16/1000
70/70 - 0s - loss: 1.1636 - accuracy: 0.5477
Epoch 17/1000
70/70 - 0s - loss: 1.1577 - accuracy: 0.5591
Epoch 18/1000
```

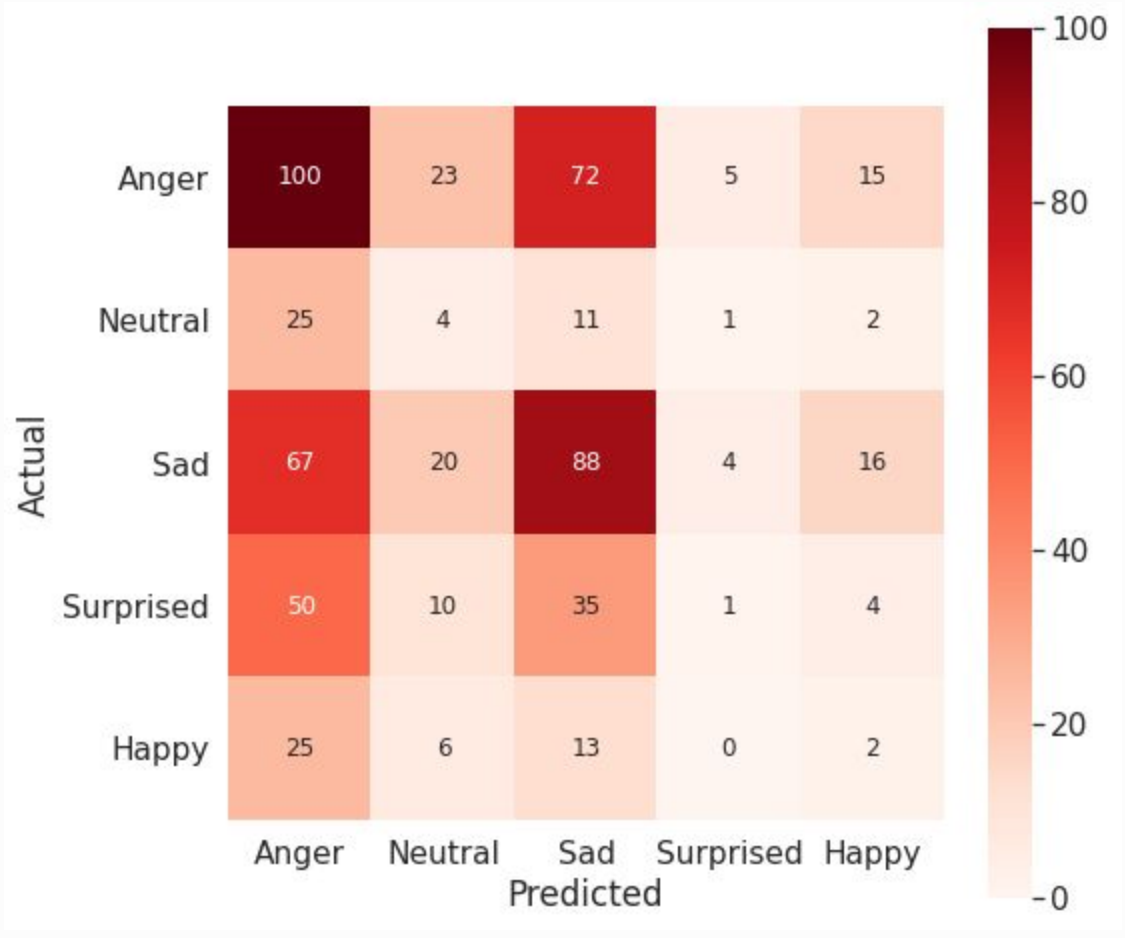
```
70/70 - 0s - loss: 0.2903 - accuracy: 0.8982
Epoch 208/1000
70/70 - 0s - loss: 0.2662 - accuracy: 0.9061
Epoch 209/1000
70/70 - 0s - loss: 0.2456 - accuracy: 0.9111
Epoch 210/1000
70/70 - 0s - loss: 0.2694 - accuracy: 0.9047 - val_loss: 2.9300 - val_accuracy: 0.3773
Epoch 211/1000
70/70 - 0s - loss: 0.2508 - accuracy: 0.9061
Epoch 212/1000
70/70 - 0s - loss: 0.2796 - accuracy: 0.8968
Epoch 213/1000
70/70 - 0s - loss: 0.2730 - accuracy: 0.9004
Epoch 214/1000
70/70 - 0s - loss: 0.2658 - accuracy: 0.9025
Epoch 215/1000
70/70 - 0s - loss: 0.2546 - accuracy: 0.9104 - val_loss: 3.2074 - val_accuracy: 0.3539
Epoch 216/1000
70/70 - 0s - loss: 0.2651 - accuracy: 0.9104
Epoch 217/1000
70/70 - 0s - loss: 0.2665 - accuracy: 0.9054
Epoch 218/1000
70/70 - 0s - loss: 0.2656 - accuracy: 0.8961
Epoch 219/1000
70/70 - 0s - loss: 0.2629 - accuracy: 0.9075
Epoch 220/1000
70/70 - 0s - loss: 0.2591 - accuracy: 0.9075 - val_loss: 3.0572 - val_accuracy: 0.3623
Epoch 221/1000
70/70 - 0s - loss: 0.3178 - accuracy: 0.8867
Epoch 222/1000
70/70 - 0s - loss: 0.2624 - accuracy: 0.9054
Epoch 223/1000
70/70 - 0s - loss: 0.2539 - accuracy: 0.9039
Epoch 224/1000
70/70 - 0s - loss: 0.3024 - accuracy: 0.8889
Epoch 00224: early stopping
```


Anger
Neutral
Sad
Surprised
Happy
Actual
76
19
92
1
27
10
4
15
1
13
71
12
88
3
21
38
12
35
2
13
20
1
17
0
8
Anger
Neutral
Sad
Surprised
Happy
Predicted
80
60
40
20
0

Mel Spectrogram on all:

```
Epoch 1/1000
70/70 - 0s - loss: 1.2288 - accuracy: 0.4925
Epoch 2/1000
70/70 - 0s - loss: 1.1847 - accuracy: 0.5133
Epoch 3/1000
70/70 - 0s - loss: 1.1627 - accuracy: 0.5219
Epoch 4/1000
70/70 - 0s - loss: 1.1305 - accuracy: 0.5326
Epoch 5/1000
70/70 - 0s - loss: 1.0948 - accuracy: 0.5505 - val_loss: 1.3545 - val_accuracy: 0.4775
Epoch 6/1000
70/70 - 0s - loss: 1.0960 - accuracy: 0.5677
Epoch 7/1000
70/70 - 0s - loss: 1.0573 - accuracy: 0.5957
Epoch 8/1000
70/70 - 0s - loss: 1.0064 - accuracy: 0.6036
Epoch 9/1000
70/70 - 0s - loss: 1.0325 - accuracy: 0.5971
Epoch 10/1000
70/70 - 0s - loss: 0.9710 - accuracy: 0.6065 - val_loss: 1.4200 - val_accuracy: 0.4674
Epoch 11/1000
70/70 - 0s - loss: 0.9358 - accuracy: 0.6323
Epoch 12/1000
70/70 - 0s - loss: 0.9093 - accuracy: 0.6581
Epoch 13/1000
70/70 - 0s - loss: 0.8704 - accuracy: 0.6659
Epoch 14/1000
70/70 - 0s - loss: 0.8362 - accuracy: 0.6839
Epoch 15/1000
70/70 - 0s - loss: 0.8408 - accuracy: 0.6746 - val_loss: 1.5370 - val_accuracy: 0.4524
Epoch 16/1000
70/70 - 0s - loss: 0.8081 - accuracy: 0.6867
Epoch 17/1000
70/70 - 0s - loss: 0.7695 - accuracy: 0.7125
```

```
Epoch 172/1000
70/70 - 0s - loss: 0.0636 - accuracy: 0.9792
Epoch 173/1000
70/70 - 0s - loss: 0.0512 - accuracy: 0.9878
Epoch 174/1000
70/70 - 0s - loss: 0.0816 - accuracy: 0.9706
Epoch 175/1000
70/70 - 0s - loss: 0.0924 - accuracy: 0.9670 - val_loss: 4.0621 - val_accuracy: 0.4374
Epoch 176/1000
70/70 - 0s - loss: 0.0820 - accuracy: 0.9713
Epoch 177/1000
70/70 - 0s - loss: 0.0716 - accuracy: 0.9778
Epoch 178/1000
70/70 - 0s - loss: 0.0635 - accuracy: 0.9792
Epoch 179/1000
70/70 - 0s - loss: 0.0688 - accuracy: 0.9771
Epoch 180/1000
70/70 - 0s - loss: 0.0651 - accuracy: 0.9806 - val_loss: 3.9980 - val_accuracy: 0.4541
Epoch 181/1000
70/70 - 0s - loss: 0.0645 - accuracy: 0.9842
Epoch 182/1000
70/70 - 0s - loss: 0.0545 - accuracy: 0.9828
Epoch 183/1000
70/70 - 0s - loss: 0.0614 - accuracy: 0.9763
Epoch 184/1000
70/70 - 0s - loss: 0.0512 - accuracy: 0.9857
Epoch 185/1000
70/70 - 0s - loss: 0.0586 - accuracy: 0.9792 - val_loss: 4.0866 - val_accuracy: 0.4357
Epoch 186/1000
70/70 - 0s - loss: 0.0667 - accuracy: 0.9828
Epoch 187/1000
70/70 - 0s - loss: 0.0680 - accuracy: 0.9806
Epoch 188/1000
70/70 - 0s - loss: 0.0592 - accuracy: 0.9806
Epoch 189/1000
70/70 - 0s - loss: 0.0851 - accuracy: 0.9749
Epoch 00189: early stopping
```



How did we choose the model?

At first, we were certain that we will need a recurrent layer because we need what happened before and our job is to spot the pattern between one sound. So There is an LSTM layer. After that we put some dense layers and that's it!

For the optimizer we tested RMSProp, Adam, default SGD from Keras and user defined SGD with lr = 0.1 and momentum = 0.9 and the best was Adam.

For the network itself we tried one or two LSTM layers, changing size of layers, and using the different activation functions such as relu, tanh and gelu! And ReLU was still the best.

Conclusion

At the end, the result was not so much desirable so there are other ideas we can implement later:

Try other features.

Try for different lengths of sounds.

Try completely different layers such as GRU or a simple RNN.

Resources:

http://ciit.finki.ukim.mk/data/papers/9CiiT/9CiiT-19.pdf

https://www.kdnuggets.com/2020/02/audio-data-analysis-deep-learning-python-part-1.html