Speech Emotion Recognition

Documentation

2CSDE70 - Natural Language Processing

18BCE046 - Yaksh Desai

18BCE054 - Dhruvil Dholariya

18BCE216 - Devam Shah

Introduction:

Recognizing emotions is one of the most important marketing strategies in today's world. You could personalize different things for an individual specifically to suit their interest. For this reason, we decided to do a project where we could recognize a person's emotions just by their voice which will let us manage many AI related applications. Some examples could be including call centers to play music when one is angry on the call. Another could be a smart car slowing down when one is angry or fearful. As a result this type of application has much potential in the world that would benefit companies and also even safety to consumers.

Dataset

For speech emotion recognition, we have used a combination of these four datasets that are mentioned below.

1. RAVDESS - Ryerson Audio-Visual Database of Emotional Speech and Song

   RAVDESS is one of the more common datasets used for this exercise by others. It's well liked because of its quality of speakers, recording and it has 24 actors of different genders.

   Modality (01 = full-AV, 02 = video-only, 03 = audio-only).

   Vocal channel (01 = speech, 02 = song).

   Emotion (01 = neutral, 02 = calm, 03 = happy, 04 = sad, 05 = angry, 06 = fearful, 07 = disgust, 08 = surprised).

   Emotional intensity (01 = normal, 02 = strong). NOTE: There is no strong intensity for the 'neutral' emotion.

   Statement (01 = "Kids are talking by the door", 02 = "Dogs are sitting by the door").

   Repetition (01 = 1st repetition, 02 = 2nd repetition).

   Actor (01 to 24. Odd numbered actors are male, even numbered actors are female). So, here's an example of an audio filename. 02-01-06-01-02-01-12.mp4

   This means the meta data for the audio file is:

   Video-only (02)

   Speech (01)

   Fearful (06)

   Normal intensity (01)

   Statement "dogs" (02)

   1st Repetition (01)

   12th Actor (12) - Female (as the actor ID number is even)

2. SAVEE - Surrey Audio-Visual Expressed Emotion

   The audio files are named in such a way that the prefix letters describes the emotion classes as follows:

   'a' = 'anger'

   'd' = 'disgust'

   'f' = 'fear'

   'h' = 'happiness'

   'n' = 'neutral'

   'sa' = 'sadness'

   'su' = 'surprise'

   The original source has 4 folders each representing a speaker, but we've bundled all of them into one single folder and thus the first 2 letter prefix of the filename represents the speaker initials. Eg. 'DC_d03.wav' is the 3rd disgust sentence uttered by the speaker DC. It's worth nothing that they are all male speakers only.

   TESS - Toronto emotional speech set

   Now on to the TESS dataset, it's worth nothing that it's only based on 2 speakers, a young female and an older female. This should hopefully balance out the male dominant speakers that we have on SAVEE.

3. CREMA-D - Crowd-sourced Emotional Multimodal Actors Dataset
   CREMA dataset, Not much is known about this dataset and we don't see much usage of this in general in the wild. But it's a very large dataset which we need. And it has a good variety of different speakers, apparently taken from movies. And the speakers are of different ethnicities. This is good. Means better generalisation when we do transfer learning. Very important

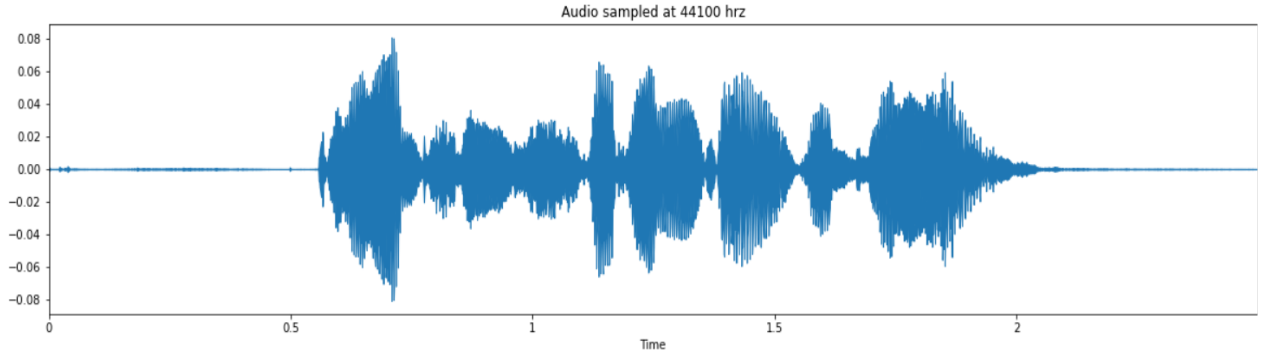
Implementation

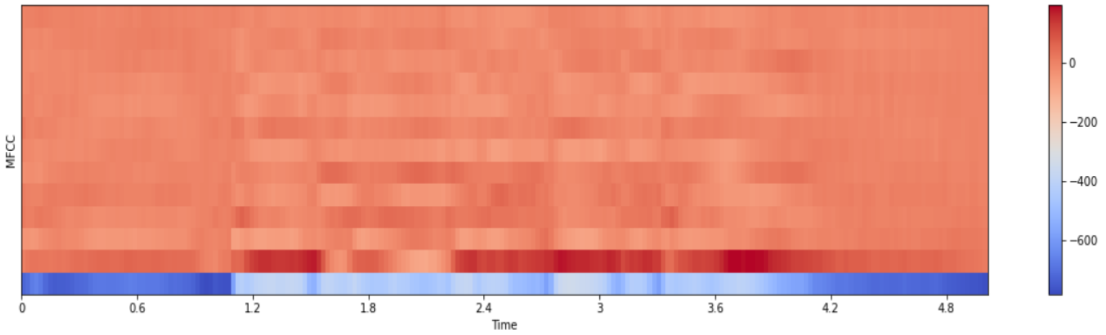
Approach 1 - 2D CNN Model with MFCC features as input

Here MFCC stands for Mel-frequency cepstral coefficient. A spectrogram is a representation of speech over time and frequency. 2D convolution filters help capture 2D feature maps in any given input. Such rich features cannot be extracted and applied when speech is converted to text and or phonemes. Spectrograms, which contain extra information not available in just text, are capable of improving emotion recognition.

The MFCC technique is the one of the most popular spectral based parameters used in recognition system approach. The MFCC has an advantage over servals techniques which is less complexity in implementation algorithms for feature extraction. The MFCC feature extraction technique basically includes several filters starting from windowing filter, Discrete Fourier Transform filter, log filter for the magnitude, etc. The MFCC extracted all the samples and then statistically analyzed the principal components, at least 2D minimally required in further recognition performance evaluation

For each audio file we have taken 30 MFCC bands of 216 audio length and given it to the 2D CNN for training purposes.

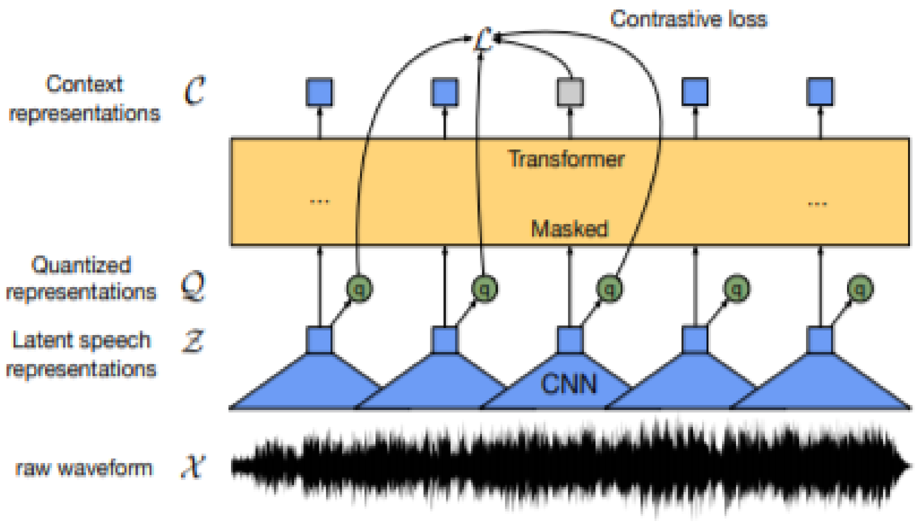




Approach 2 - Emotion Recognition from Speech Using Wav2vec 2.0 Embeddings

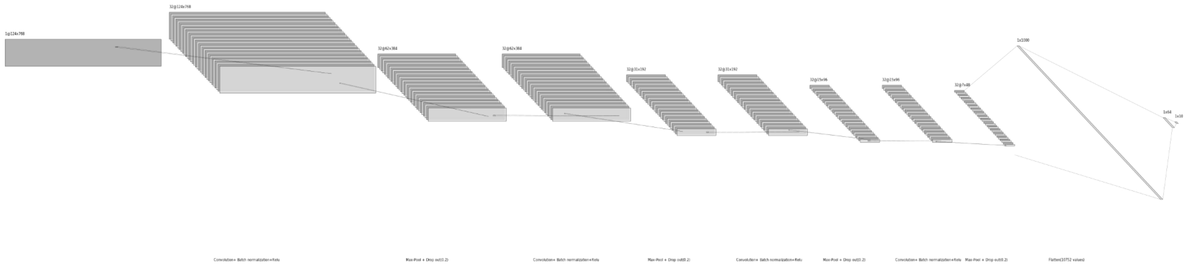
Wav2vec 2.0 is a framework for self-supervised learning of representations from raw audio. The model consists of three stages. The first stage is a local encoder, which contains several convolutional blocks and encodes the raw audio into a sequence of embeddings with a stride of 20 ms and a receptive field of 25 ms. Two models have been released for public use, a large one and a base one where the embeddings are 1024- and 768-dimensional, respectively. The second stage is a contextualized encoder, which takes the local encoder representations as input. Its architecture consists of several transformer encoder blocks. The base model uses 12 transformer blocks with 8 attention heads each, while the large model uses 24 transformer blocks with 16

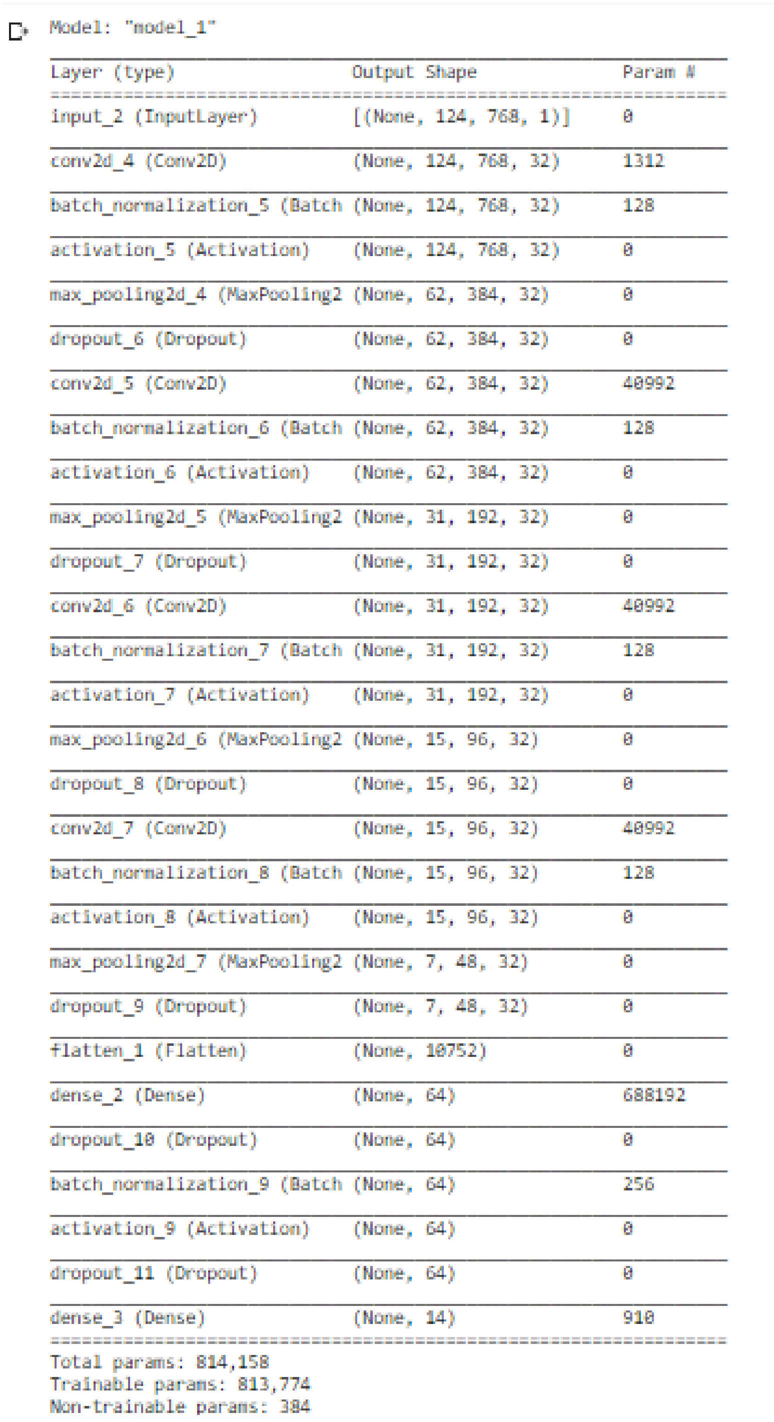
attention heads each. Finally, a quantization module, takes the local encoder representations as input and consists of 2 codebooks with 320 entries each.
We have used features as the weighted average of the outputs of the 12 transformer blocks. The weights of this average are trained along with the 2D CNN model.



Architecture of 2D CNN

Input image will be of size (30,216) in the case of the 1st approach while in the second approach it will be a (124,768) size numpy array. Then in both cases we have applied 4 CNN blocks of the same architecture. In this block the first Conv2D of 32 filters having kernel size=(4,10) is applied while the padding value is set as the same. So there will be no change in shape. Then Batch Normalization and Relu activation is applied on the same. Then MaxPooling of (2,2) is applied so the new shape will be half of the old one and then drop out of 0.2. After these four blocks the final shape will be (7,48,32) and then it has got flattened and it will be of (10752). Then there is a dense layer of 64 neurons followed by drop out, batch normalization and relu activation and then the final output layer will be of 14 neurons with softmax activation.



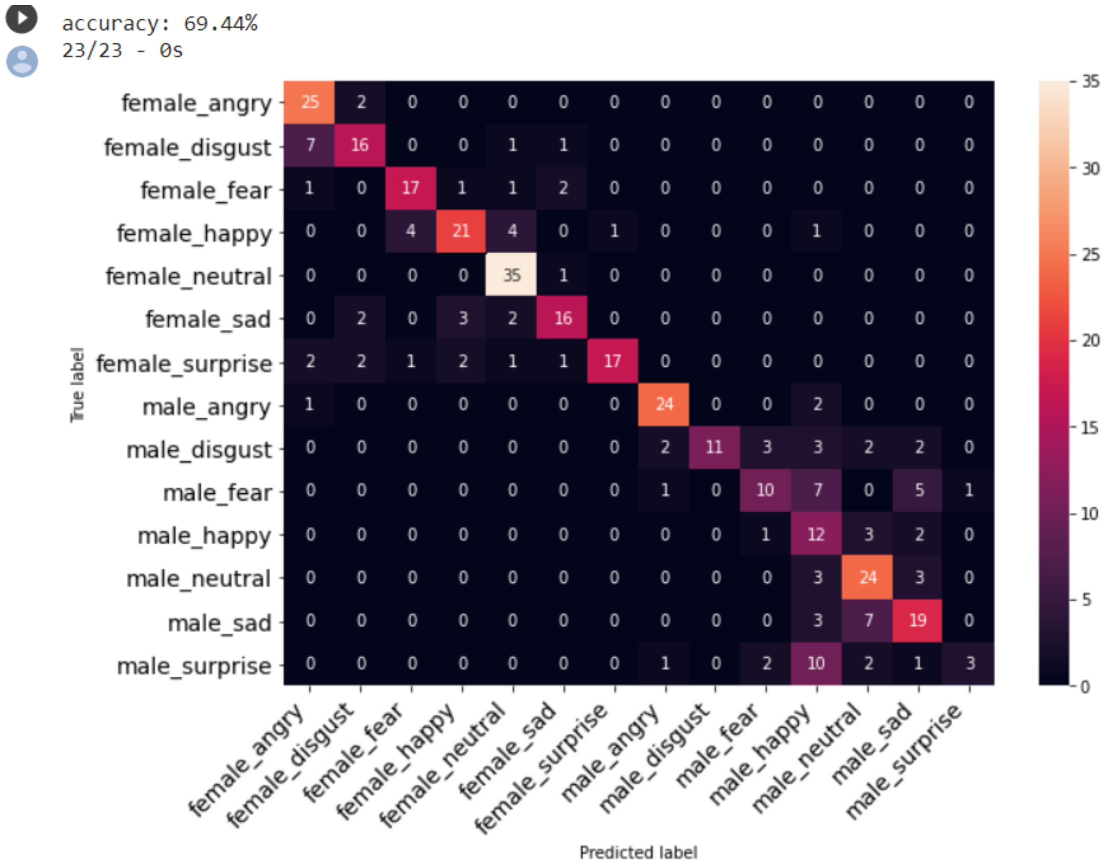


```
Model: "model_1"
_________________________________________________________________
Layer (type)                 Output Shape              Param #   
=================================================================
input_2 (InputLayer)         [(None, 124, 768, 1)]     0         
_________________________________________________________________
conv2d_4 (Conv2D)            (None, 124, 768, 32)      1312      
_________________________________________________________________
batch_normalization_5 (Batch (None, 124, 768, 32)      128       
_________________________________________________________________
activation_5 (Activation)    (None, 124, 768, 32)      0         
_________________________________________________________________
max_pooling2d_4 (MaxPooling2 (None, 62, 384, 32)       0         
_________________________________________________________________
dropout_6 (Dropout)          (None, 62, 384, 32)       0         
_________________________________________________________________
conv2d_5 (Conv2D)            (None, 62, 384, 32)       40992     
_________________________________________________________________
batch_normalization_6 (Batch (None, 62, 384, 32)       128       
_________________________________________________________________
activation_6 (Activation)    (None, 62, 384, 32)       0         
_________________________________________________________________
max_pooling2d_5 (MaxPooling2 (None, 31, 192, 32)       0         
_________________________________________________________________
dropout_7 (Dropout)          (None, 31, 192, 32)       0         
_________________________________________________________________
conv2d_6 (Conv2D)            (None, 31, 192, 32)       40992     
_________________________________________________________________
batch_normalization_7 (Batch (None, 31, 192, 32)       128       
_________________________________________________________________
activation_7 (Activation)    (None, 31, 192, 32)       0         
_________________________________________________________________
max_pooling2d_6 (MaxPooling2 (None, 15, 96, 32)        0         
_________________________________________________________________
dropout_8 (Dropout)          (None, 15, 96, 32)        0         
_________________________________________________________________
conv2d_7 (Conv2D)            (None, 15, 96, 32)        40992     
_________________________________________________________________
batch_normalization_8 (Batch (None, 15, 96, 32)        128       
_________________________________________________________________
activation_8 (Activation)    (None, 15, 96, 32)        0         
_________________________________________________________________
max_pooling2d_7 (MaxPooling2 (None, 7, 48, 32)         0         
_________________________________________________________________
dropout_9 (Dropout)          (None, 7, 48, 32)         0         
_________________________________________________________________
flatten_1 (Flatten)          (None, 10752)             0         
_________________________________________________________________
dense_2 (Dense)              (None, 64)                688192    
_________________________________________________________________
dropout_10 (Dropout)         (None, 64)                0         
_________________________________________________________________
batch_normalization_9 (Batch (None, 64)                256       
_________________________________________________________________
activation_9 (Activation)    (None, 64)                0         
_________________________________________________________________
dropout_11 (Dropout)         (None, 64)                0         
_________________________________________________________________
dense_3 (Dense)              (None, 14)                910       
=================================================================
Total params: 814,158
Trainable params: 813,774
Non-trainable params: 384
```

Results

Approach	Accuracy on RAVDESS, TESS, SAVEE, CREMA-D
2D CNN Model with MFCC features as input	69.44%
Emotion Recognition from Speech Using Wav2vec 2.0 Embeddings	63.11%

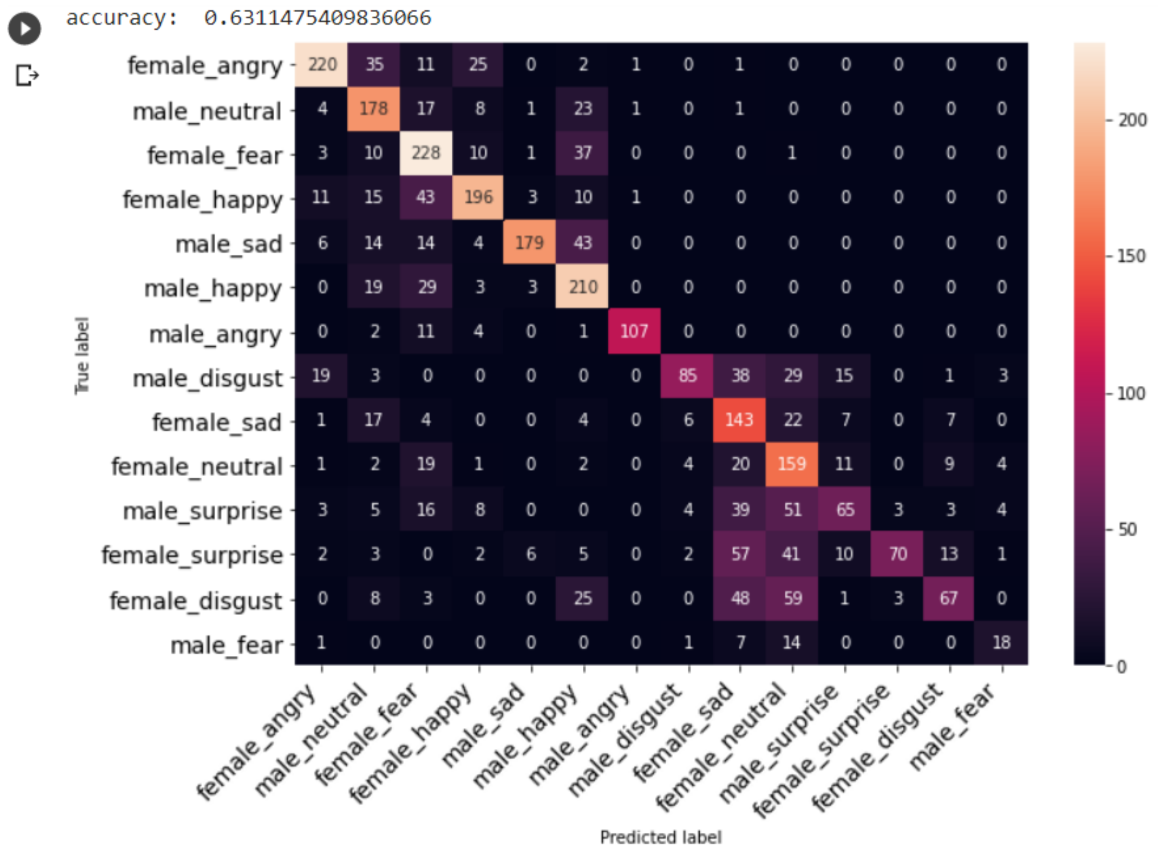
Approach 1



```
accuracy: 69.44%
23/23 - 0s
```

	female_angry	female_disgust	female_fear	female_happy	female_neutral	female_sad	female_surprise	male_angry	male_disgust	male_fear	male_happy	male_neutral	male_sad	male_surprise
female_angry	25	2	0	0	0	0	0	0	0	0	0	0	0	0
female_disgust	7	16	0	0	1	1	0	0	0	0	0	0	0	0
female_fear	1	0	17	1	1	2	0	0	0	0	0	0	0	0
female_happy	0	0	4	21	4	0	1	0	0	0	1	0	0	0
female_neutral	0	0	0	0	35	1	0	0	0	0	0	0	0	0
female_sad	0	2	0	3	2	16	0	0	0	0	0	0	0	0
female_surprise	2	2	1	2	1	1	17	0	0	0	0	0	0	0
male_angry	1	0	0	0	0	0	0	24	0	0	2	0	0	0
male_disgust	0	0	0	0	0	0	0	2	11	3	3	2	2	0
male_fear	0	0	0	0	0	0	0	1	0	10	7	0	5	1
male_happy	0	0	0	0	0	0	0	0	0	1	12	3	2	0
male_neutral	0	0	0	0	0	0	0	0	0	0	3	24	3	0
male_sad	0	0	0	0	0	0	0	0	0	0	3	7	19	0
male_surprise	0	0	0	0	0	0	0	1	0	2	10	2	1	3

Approach 2



```
accuracy:  0.6311475409836066
```

True label \ Predicted label	female_angry	male_neutral	female_fear	female_happy	male_sad	male_happy	male_angry	male_disgust	female_sad	female_neutral	male_surprise	female_surprise	female_disgust	male_fear
female_angry	220	35	11	25	0	2	1	0	1	0	0	0	0	0
male_neutral	4	178	17	8	1	23	1	0	1	0	0	0	0	0
female_fear	3	10	228	10	1	37	0	0	0	1	0	0	0	0
female_happy	11	15	43	196	3	10	1	0	0	0	0	0	0	0
male_sad	6	14	14	4	179	43	0	0	0	0	0	0	0	0
male_happy	0	19	29	3	3	210	0	0	0	0	0	0	0	0
male_angry	0	2	11	4	0	1	107	0	0	0	0	0	0	0
male_disgust	19	3	0	0	0	0	0	85	38	29	15	0	1	3
female_sad	1	17	4	0	0	4	0	6	143	22	7	0	7	0
female_neutral	1	2	19	1	0	2	0	4	20	159	11	0	9	4
male_surprise	3	5	16	8	0	0	0	4	39	51	65	3	3	4
female_surprise	2	3	0	2	6	5	0	2	57	41	10	70	13	1
female_disgust	0	8	3	0	0	25	0	0	48	59	1	3	67	0
male_fear	1	0	0	0	0	0	0	1	7	14	0	0	0	18