Speech Emotional Recognition

Aaparna Balan - S1042839

I. INTRODUCTION

Speech Emotion Recognition (SER) has gained a lot of attention in understanding the human emotional behaviors [1]. The importance of speech is understood when we use different mode of communication such as (texts or emails.)[2]. As we realize that emotion detection of speech helps to understand each other better, it has also extended the outcomes for the computers to understand emotions. Almost all the systems recognizes and detects speech signals in day today life. The SER can be used in wide range of application such as call centers, robots, computer games etc.

In speech recognition there exists two important types of speech information

- Linguistic information (context of the speech)
- Paralinguistic information

The paralinguistic information targets the emotions portrayed in the speech [3]. The speech emotion recognition can be detected and distinguished into different areas as shown in Fig.1. The SER system collects all the inputs of the speaker and uses different methodologies to process those input signals for detecting the emotions. The speech can either be natural,acted, or elicited [7]. The use of optimal features (prosody, spectral,voice quality,etc) in speech emotion recognition helps to capture the nuances of the speech utterances and are used while training the model. The resultant emotions can be and are classified as positive,negative and neutral.

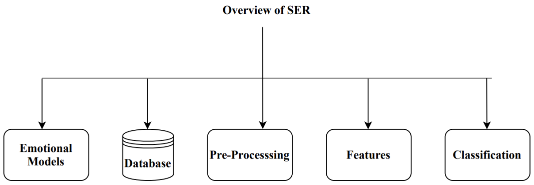


Fig. 1. Overview of SER

A. Problem

The paper mainly focuses on understanding the speech signals and predicts the emotions for different data sets which is explained in section III. It is difficult to distinguish the emotions, for instance the expression of calm vary from person to person which is more likely to be neutral in few cases. These utterances makes the model to mispredict resulting in less accuracy score. Expressions are deep and it will be easy for a human to understand the sense or tone of the speech while communicating, but the model has to capture these essence of tonality in order to perform better. Though naturally spoken recordings will equally carry the context as that of the enacted recordings but it is arduous to produce outstanding results.

B. Background

There are lot of related publications on speech emotion recognition and focuses on different areas of interpretations. The paper by Schuller et al. [4] focuses on the performance of the classifier by using differences in noise level (miss match in noise level, change in microphone settings, speaker variation, and so on). They showed potential improvement in the results can be achieved by using optimized features for these differences in noise. Further Mohammed Abdelwahab and Carlos Busso [5] explored a supervised method to hike the performance of the system by evaluating the system with a mismatched training and test conditions. It is also important to differentiate the neutral speech especially while expressing emotions. An iterative feature normalization (IFN) algorithm detected the neutral speech by using normalization parameters [6]. This helped to classify the individual user.

Zhengwei Huang[1], proposed a CNN model which facilitated categorization for the hierarchical input data. The candidate set were trained to learn the affect-salient and discriminate features for which the model performance was robust. The choice of right feature for the model will boost its performance and Oh-Wook Kwon [8] proved that pitch and energy are the salient features for emotion recognition. The utterance level emotions were recognized and obtained 20% of accuracy to that of the traditional state-of-the-art model [9]. A CNN method was introduced by W. Q. Zheng et al which used labeled audio data and outperformed SVM classifier [10].

C. Research Question and Hypotheses

The speech emotion recognizer helped to detect the expressions of the speakers by extracting the important feature and using it in the model. To further improve and analyse the performance of the classifier two concepts have been addressed for which the performance is evaluated.

- How different is it from SER classifier to that of the human annotation for the emotions? What percent of label data is needed to improve performance?
- Which is better in terms of performance? Naturally spoken data or spontaneously received information?

OVERVIEW

The paper uses CNN model to captures emotions for different data sets. By using data augmentation the difference between clear speech and noisy data can be distinguished. The above mentioned research questions are addressed by training the model with natural and enacted interactions. Apart from that the human annotated data is compared with the model prediction to gauge the performance and to improve the accuracy.The MFCC feature is used along with other salient features discussed in section III. The results produced by the model is analyzed which also uses spontaneous data set to validate and is discussed in further sections.

II. SET-UP

A. Data Exploration

The study relies on different data set which are available to train the emotional classifier. The RAVDESS and CREMA-D [11] is used for training and used Common Voice data to test the two model.

RAVDESS: It is a multimodal database of emotional speech [12]. It is a gender balanced data set with a North American accent. It comprises of 1440 files, seven expressions along with the intensity of the emotions.The purpose of using this data set is to understand the emotional behaviors.

CREMA-D: The data is diverse with different accents from 91 actors. It hold 7442 recordings which is also gender balanced and the sentences express the emotions. This database allows to train the model more efficiently due to diversity of the speaker.

SAVEE: The SAVEE database was recorded from four native English male speakers .Emotion has been described psychologically in discrete categories: anger, disgust, fear, happiness, sadness and surprise. A neutral category is also added to provide recordings of 7 emotion categories [13].

Common Voice: It consists data read by the users from public domains. It included speakers of different age groups with different accent and not gender biased. These enacted data is used to test the developed model. It is also interesting to determine the importance of tonality, diversity and spontaneity with accordance to the emotions.

Each of these recordings are annotated and compared with the predictions to measure the model performance. Various dimensions of the data is exposed and the most salient features are used for training.

III. METHOD

A. Methodology

The speech emotion recognition uses different data set and it extracts important features to preserve the variation of the speakers and recordings while processing it. The data is combined in such a way that the gender based emotions are obtained. This helps us to easily distinguish the gender and its associated emotions.

Silence removal: The silence signal in the speech is not informative and it is removed by computing the maximum value of the frame by comparing it to a threshold value.

Pre-emphasis: The frequency of the input data vary from speaker to speaker and to make it comparable (all format) the signals are passed to a high pass filter which can be applied and the (pre-emphasis) parameter is taken as 0.93 [14].

$$x'(n) = x(n) - x(n-1) \tag{1}$$

Data Normalization: The volume of the speech sounds has to be changed to a normalized level for which the sequence is divided by highest value to receive a similar value for the entire input signals.

Windowing: The pauses or stopping places in the audio signals can be reduces and hamming window is used to do that and is represented in Eq.2 [14]

$$w(n) = 0.54 - 0.46cos(2 * pi * n)/(M-1) \tag{2}$$

Data Augmentation

In data augmentation small perturbations are added to the new synthetic data on the initial training set [15]. The adaptation of neural network for speech recognition and to train the models the data augmentation is very helpful and indeed, improves accuracy.

Pitch: The change in pitch raises or lowers the input waves in certain intervals. The librosa package in python makes the task easier for which the pitch of the audio signals are changed and used while training. The output signals after this augmentation seems to be distinct and the variation in the tone can be established.

Stretch: The input waves are expanded and each word pronunciation becomes better.

Speed and pitch: The pitch is changed uniformly with the increase in speed. This can also have a variation in pitch by changing the low or high values with the change in speed of the audio. It is one of the salient method for our model.

Noise: The size of the training data set expands while adding noise and it also help to avoid overfitting. This white noise in the data has equal intensity and produces a masking effect. *Dynamic change*: The prosodic information is carried out in the dynamic change which easily facilitates to convey the emotions by the speaker [16]. The received input signals must be processed and made model ready in order to receive the best accuracy.

Data Visualisation

The total percentage of emotions for the above mentioned data set is show in fig.2.

The loudness of the audio can be known through the Waveplots and the Spectrum frequency are visualized using Spectograms with respect to time as shown in Fig.3.

Feature Extraction

The audio features can be categorised to time and frequency domain where the time domain features vary based on time and includes few other features such as minimum energy, the maximum excursions. On the other hand the frequency domain used the signals in the form of frequency and not time. This helps to capture the emotions from the speech signals as the time and frequency domains provide a great insight in recognizing the input signals. The pitch can be considered as a main vocal signal for speech emotion recognition.

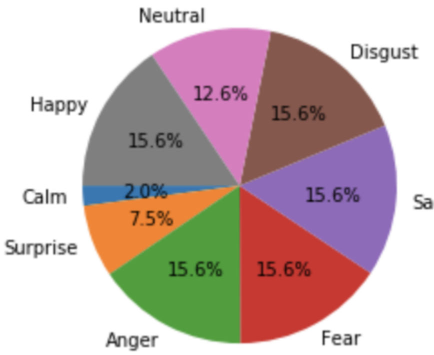


Fig. 2. Count for each emotions

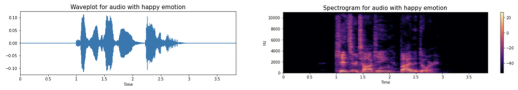


Fig. 3. Waveplot and Spectogram for happy emotion

To extract the salient features for our model the Mel-Frequency Cepstral Coefficients (MFCCs), spectral centroid, spectograms, delta-delta feature were studies and chose the two major important features, (i). Mel-spectogram and (ii), Mel-Frequency Cepstral Coefficients (MFCCs)

Log Mel-spectogram

In Mel-spectogram, the mel-scale is obtained by converting the frequencies. The pattern received from the speech audio emotion varies based on the expression in the energy spectrum. The inputs are sampled to the window size and the hop length provides the size for the hop to the next window. The log in Mel-spectogram allows us to understand the perceived emotions from the recording for different datasets.

Mel-Frequency Cepstral Coefficients (MFCCs)

The MFCC takes the speakers audio signals and analyses the emotions based on the frequencies, converts these frequencies to Mel scale which preserves the actual tone of the audio. To compute MFCCs, we have to first frame the audio signals to multiple smaller frames. Next, the periodogram is calculated for each frame which detects the frequencies in these sub frames. Now to understand the energy involved in each frames the Mel-filter is used which adds up the energy in each small frames. As discussed in the log Mel-spectogram, the speech signals dorsnot have a linear frequency and hence to capture that we need a log scale (applied to the filterbank). To avoid correlation with the frequency band a discrete cosine transform (DCT) is used and the coefficients are maintained between 2-13, discarding the rest [17].

These two features provides a good improvement in CNN model and helps to classify the emotions.

B. *Selection of method*

The methods proposed in the paper is biased in various aspects. This section explains the selection of the methods and the hidden limitations for which adaptations of various methodologies is discussed in the section VI.
Firstly, the paper used three different datasets (RAVDESS and CREMA and SAVEE). These datasets were trained to obtain a better performing model but it is bias in terms of accent, dialects, gender. The model is also tested on the spontaneous utterance were the performance of the model delineates for naturally spoken data. The hidden difficulty is that the model is trained to perform good for English language but the utterances of other languages will lead to mispredictions.
Secondly, understand the emotional patterns based on the linguistic elements is difficult. The influence of diverse data (CREMA) helps to boost the model.

IV. Experiment

Data augmentation helps to improve diversity in the training data for which I will explain the two main augmented methods. The white noise is induced in the training data which will detect the emotions of the speakers and improves diversity in the data.The dynamic change and speed and pitch improves the data with a variation in the frequency. The pitch of each speaker varies, but by augmenting it becomes distinct and the data used while training can be easily identified and reduced false negative improving recall. As discussed earlier MFCCs were extracted, and from each frame 13 features were returned which is shown in the table.1. The log mel-spectogram features were also used with a window size of 0.0018 and hop of 0.005 sec. The pitch, augmented noise and dynamic change was added as features along with MFCCs.

A. *Parameters*

Parameter |	|Types
Re_sampling	kaiser_fast
Duration	2.5
Sampling_rate	45100
Offset	0.5
n$_{m}fcc$	13
Day	Min Temp

TABLE I
PARAMETER TYPE

B. *Train and Test Data*

To tune the model few steps are established for improving the performance.The dataset comprises different speakers expressing their emotions and same speakers portraying different moods.To avoid over-fitting few data from CREMA (25%) is used as validate set in the model. This is because it comprises of data which is diverse, uses different actors and is naturally spoken. Later the model is also tested by using Common voice database. The main purpose of doing such tests is to show that the model is not gender biased and results for predicting emotions of different age groups. In addition, the model is also not trained with limited actors which will not result in poor performance while using Common voice data.

C. *MODEL*

Convolution Neural Networks

The speech emotion recognition can be performed in various models and the paper uses CNN for training the raw audio signals, classifying and testing it with the use of GPU.

Why CNN?

The CNN automatically detected the important features and can be used for large dataset in speech recognition and are computational efficient. As CNN already is a good feature extractor while combining it with MFCC, introduces a linearity in the log-frequency axis [18]. This makes it easier to adapt the features and understand the frequencies and patters in the frames resulting a good performance.

A 1D - Convolution network is used, which uses sequential model. The Keras library in python allows us to define the model, were each layer will have only one input and output tensor. Here 8 1D-layers are used followed by one dense layer. Followed by the convolution layer is the activation function. The main purpose for using Relu is that it reduces the vanishing gradient problems and the weights are biased for some neurons during back propagation (not all neurons are active) allowing the model to learn faster. The batch normalization helps to improve the learning rates with faster execution while training. The important features must be extracted in order to enhance the model performance were max pooling captures even the basic features.

Finally, all the neurons are connected to each other which is defined by the use of dense layer in the model. To fit the model, a batch size of 15 was used with a total of 150 epochs.

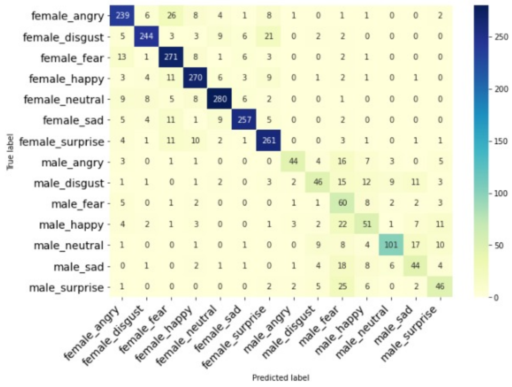
The model predicted all the emotions (happy,sad,anger,frustrated,disgust,calm,neutral, surprised) and accuracy is used as the evaluation metrics for the emotion detection from the input signals. The model was also tested using a different data set which is highly diverse than that of the training set. It is evident that the hidden problems makes the model bias in certain way, but as the model was also trained using CREMA dataset, its perform was seemingly good.

I also compared the model performance with a base model, which used only RAVDESS and CREMA data with a train test split of 80% and 20%. The performance is not good and the prediction results are poor . The one reason for this is that the data is not diverse .

V. Analysis and results

The different featuring methods and modeling was used to predict the speech emotions. As the MFCC and log mel-spectogram was used the model performed pretty good with an overall accuracy of 79%. A confusion matrix is used to show the model performance Fig.4.

Now lets do some comparison with the obtained results to gauge the usefulness of the model as the main goal is to focus on the following questions.



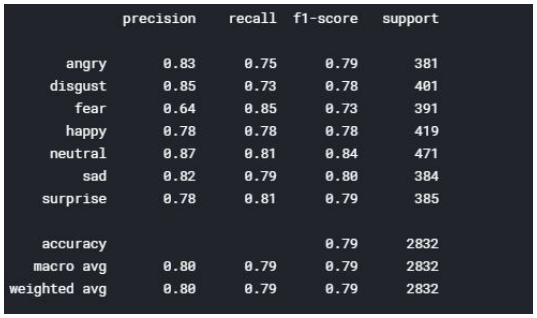
	female_angry	female_disgust	female_fear	female_happy	female_neutral	female_sad	female_surprise	male_angry	male_disgust	male_fear	male_happy	male_neutral	male_sad	male_surprise
female_angry	239	6	26	8	4	1	8	1	0	1	1	0	0	2
female_disgust	5	244	3	3	9	6	21	0	2	2	0	0	0	0
female_fear	13	1	271	8	1	6	3	0	0	2	1	0	0	0
female_happy	3	4	11	270	6	3	9	0	1	2	1	0	1	0
female_neutral	9	8	5	8	280	6	2	0	0	1	0	0	0	0
female_sad	5	4	11	1	9	257	5	0	0	2	0	0	0	0
female_surprise	4	1	11	10	2	1	261	0	0	3	1	0	1	1
male_angry	3	0	1	1	0	0	0	44	4	16	7	3	0	5
male_disgust	1	1	0	1	2	0	3	2	46	15	12	9	11	3
male_fear	5	0	1	2	0	0	0	1	1	60	8	2	2	3
male_happy	4	2	1	3	0	0	1	3	2	22	51	1	7	11
male_neutral	1	0	0	1	0	1	0	0	9	8	4	101	17	10
male_sad	0	1	0	2	1	1	0	1	4	18	8	6	44	4
male_surprise	1	0	0	0	0	0	2	2	5	25	6	0	2	46

Fig. 4. Confusion Matrix

How different is it from SER classifier to that of the human annotation for the emotions? What percent of label data is needed to improve performance?

The data set was annotated by the human and compared with the produced results. As a result, the human annotated data and the model predicted results are more or less similar where the human had much more clarity in understanding the disgust and happy emotion.Humans tend to perceive emotions at a high rate as communication is the source of understanding emotions. The system used augmented data and features to predict the emotions and captured most of the speakers emotions. The nuances are always difficult to detect and the hidden conditions can also reflect it the predictions for which it resulted a bit less for fear and disgust. Overall the model performed better equal to the human annotated data but failed in capturing few emotions. These are the top four emotions of the annotated data : 95.4 % happy, 89 % angry,93.4% sad, and 86.4% fear.

It is also evident that the model performance can be improved with an improvement of 8 to 15% in adaptation of label data .Fig.6.



```
              precision    recall  f1-score   support

       angry       0.83      0.75      0.79       381
     disgust       0.85      0.73      0.78       401
        fear       0.64      0.85      0.73       391
       happy       0.78      0.78      0.78       419
     neutral       0.87      0.81      0.84       471
         sad       0.82      0.79      0.80       384
    surprise       0.78      0.81      0.79       385

    accuracy                           0.79      2832
   macro avg       0.80      0.79      0.79      2832
weighted avg       0.80      0.79      0.79      2832
```

Fig. 5. Model Predictions

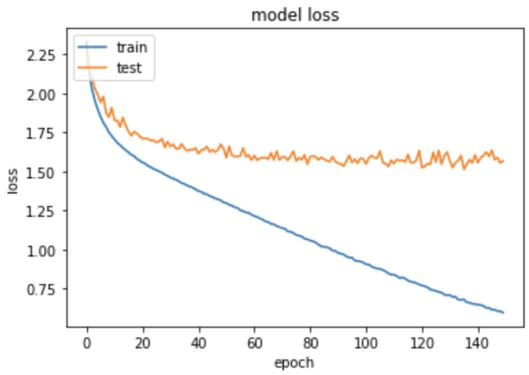
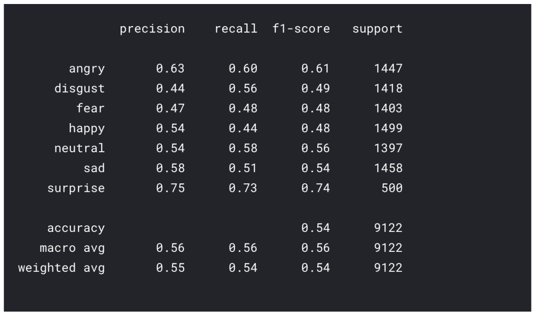


Fig. 6. Model loss

I would also like to address that the data used are gender balanced and the proposed results are not gender biased and I am not using gender for the above comparison. Gender can lead to various dimensions and analysis which can be used for future works.

Which is better in terms of performance? Naturally spoken data or spontaneously received information?
As the model was trained based on both enacted (40%) as well as naturally spoken data. The result of the initial case using natural data are not satisfactory Fig.6. This base model was discussed early and it was compared with the final CNN model testing upon spontaneous data. The result is evident that the spontaneous acted recordings are more useful than the naturally spoken inputs.
The base model used RAVDESS(enacted) and natural data for which the features were extracted and MFCC was used. The overall accuracy is 54% and the model performs poor for fear (48%) and happy(48%). On the other hand, as discussed in experiment session the final CNN used log Mel-spectogram and other features which also validated using CREMA data set. The result boosted up will a less loss function achieving an accuracy of 73% and 78% for happy and fear emotions. This clearly supports the hypothesis that the spontaneous data is importance to improve the score.


```
              precision    recall  f1-score   support

       angry       0.63      0.60      0.61      1447
     disgust       0.44      0.56      0.49      1418
        fear       0.47      0.48      0.48      1403
       happy       0.54      0.44      0.48      1499
     neutral       0.54      0.58      0.56      1397
         sad       0.58      0.51      0.54      1458
    surprise       0.75      0.73      0.74       500

    accuracy                           0.54      9122
   macro avg       0.56      0.56      0.56      9122
weighted avg       0.55      0.54      0.54      9122
```

Fig. 7. Model Predictions for naturally used data

VI. Discussion

As humans can easily communicate and understand the emotions,it is most likely to achieve a good results with human annotation. But overall, the model performed good in comparison to that of the human annotated results. It is also evident that the spontaneous data is also necessary while building the model. I have not performed any analysis on gender, it might be interesting to consider gender and check the performance by training the model only with natural spoken data and test on spontaneous data.

VII. Conclusion

A CNN model was proposed to predict the speech emotion recognition. The base model comprised of data set which was trained and tested on natural occurrence of speech data and compared with the final model which was tested on spontaneous speech signals. The model performed better with an accuracy of 79% with the enacted set .
It is also surprising to see that the model performed good with comparison to the human annotated results. The robustness of the model can be achieved by training the model with both spontaneous and natural data, as in most of the cases the model performed better for enacted set.

References

[1] Zhengwei Huang,Ming Dong,Qirong Mao, Yongzhao Zhan,“Speech Emotion Recognition Using CNN,” School of computer science and Engineering, retrieved from, https://www.semanticscholar.org/paper/Speech-Emotion-Recognition-Using-CNN-Huang-Dong/6060612f69d777760e0538b8bad28b4b7607745d?citationIntent=result#citing-papers

[2] Farbod Razzazi,Saeed Dabbaghchian, Robust Speech Recognition Using MLP Neural Network in Log-Spectral Domain Conference Paper ,January 2010, retrieved from https://www.researchgate.net/publication/224111961_Robust_Speech_Recognition_Using_MLP_Neural_Network_in_Log-Spectral_Domain

[3] Jianfeng Zhaoa,b, Xia Maoa, Lijiang Chena,Speech emotion recognition using deep 1D 2D CNN LSTM networks

[4] B. Schuller, D. Seppi, A. Batliner, A. Maier, and S. Steidl, “Towards more reality in the recognition of emotional speech,” in International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2007), Honolulu, HI, USA, April 2007, vol. 4, pp. 941–944.

[5] Mohammed Abdelwahab and Carlos Busso,SUPERVISED DOMAIN ADAPTATION FOR EMOTION RECOGNITION FROM SPEECH,Multimodal Signal Processing (MSP) Laboratory, Department of Electrical Engineering The University of Texas at Dallas, Richardson TX 75080, USA

[6] C. Busso, S. Mariooryad, A. Metallinou, and S. Narayanan, “Iterative feature normalization scheme for automatic emotion detection from speech,” IEEE Transactions on Affective Com- puting, vol. 4, no. 4, pp. 386–397, October-December 2013.

[7] Mehmet Berkehan Akçaya, Kaya Oğuzb, Speech emotion recognition: Emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers

[8] Oh-Wook Kwon, Kwokleung Chan, Jiucang Hao, Te-Won Lee, Emotion Recognition by Speech Signals, Institute for Neural Computation University of California, San Diego, USA

[9] L. Chen, X. Mao, H. Yan, Text-independent phoneme segmentation combining EGG and speech data, IEEE/ACM Trans. Audio Speech Lang. Process. 24 (6)(2016) 1029–1037.

[10] W.Q. Zheng, J.S. Yu, Y.X. Zou, An experimental study of speech emotion recognition based on deep convolutional neural networks, in: International Conference on Affective Computing and Intelligent Interaction IEEE, 2015, pp. 827–831.

[11] RAVDESS Emotional speech audio,retrieved from https://www.kaggle.com/uwrfkaggler/ravdess-emotional-speech-audio

[12] The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English, retrieved from https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0196391

[13] Surrey Audio-Visual Expressed Emotion (SAVEE),retrieved from https://www.kaggle.com/ejlok1/surrey-audiovisual-expressed-emotion-savee

[14] Bashar M. Nema, Ahmed A. Abdul-Kareem, Preprocessing Signal for Speech Emotion Recognition, Department of Computer Science, College of Science, Mustansiriyah University, IRAQ

[15] Speech Emotion Recognition, retrieved from https://www.kaggle.com/shivamburnwal/speech-emotion-recognition/notebook

[16] Jing Shen1,and Pamela E. Souza, On Dynamic Pitch Benefit for Speech Recognition in Speech Masker, retrieved from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6204388/

[17] Mel Frequency Cepstral Coefficient (MFCC) tutorial, retrieved fromhttp://practicalcryptography.com/miscellaneous/machine-learning/guide-mel-frequency-cepstral-coefficients-mfccs/

[18] Kannan Venkataramanan,Haresh Rengaraj Rajamohan, Emotion Recognition from Speech, retrieved from https://arxiv.org/abs/1912.10458