Depression analysis of voice samples using machine learning

Yashi Sharma [1], Dr. Brajesh Kumar Singh [2]

[1] M. Tech. Scholar, [2] Professor, C.S.E. Department, R.B.S. Engineering Technical Campus, Bichpuri, Agra, U.P., India

ABSTRACT

Depression is seen as an emerging mental challenge in the lives of various people. Nowadays it is also becoming one of the major reasons for mental disability across the world. Depression has manifested itself as a silent killer and according to statistics it has affected more than 300 million people in United States of America majorly affecting individuals in the age group of 15 to 44 yrs. According to a study by World Health Organization, the effects of depression have been dangerous in life, it is seen causing threatening diseases like cancer, diabetic issues or even heart disease.

However, the problem that mainly is associated with the disease of depression is that it is not treated as a disease. Where the common understanding of the word “Disease” is any medical ailment that require doctor’s attention or quick medical response, depression on the other hand, even after qualifying as a disease is hidden in societal barriers to appear for a proper treatment. People whose lifestyle pattern has been intruded by depression either do not avail proper medical attention or are too shy to appear in the masses for proper attention on their physical as well as condition.

Our motivation here is to investigate through the phenomenon of depression and predict whether an individual is having symptoms of depression by accessing his/her voice sample. In order to establish a link between depression and voice features, we obtain a large data set and then train a model accordingly by applying machine learning methods on it.

This model when given a voice sample can now predict, whether a particular subject is depressed or not, to a nearby accurate measure.

KEYWORDS

Depression, machine learning, CNN, MFCC and voice sample.

INTRODUCTION

An increasing rate of depression has been seen intruding the lives of people in the present times. Lifestyle changes are creeping into our daily schedules bringing fatigue, loss of interest in day-to-day activities, personal or professional loss of commitments and many other factors which pave the way for a person being closer to depression. In most of the cases, we don’t even realize the fact that we are depressed in some way or the other until and unless we see this disease becoming dominant in our daily behavior.

The biggest drawback however with depression is the shortage to access its treatment. This shortage can be due to cost of treatment, fear of coming out in the open and being vocal about it and most importantly the lack of proper counselling. Across the globe, more than 300 million are affected by depression which brings about change of sleep patterns, varying appetite schedules, feeling low on energy and sometimes even suicidal tendencies [1,2,3]. Facts also state that 1.4% deaths in the world are due to depression and by 2030 depression will be one of the major diseases affecting lifestyle [4][5].

World Health Organization (WHO) states that the availability of a tool or method which projects accurate and reliable accessing of depressive is extremely essential. This screening of a depressive patient usually involves a questionnaire which contains varying questions whose final score or outcome would somewhat present a picture on the depressive state of the person [6].

Various researches have shown coordination between depression and a person's behavior. A primary feature of depression has been seen associated with basal ganglia where there is a lack of muscular coordination developed in the body. Research has shown that voice samples of people contain information about their mental state and can be extremely crucial in depicting the mental state of a person [7]. It has also been argued that combination of features when optimized can give accurate results for predicting depression.

Mel Frequency Cepstral Coefficients (MFCC) is the most used feature for processing speech signal as they are consistent and work well even for low dimensions [8]. Also, various research has established that when a model is trained through Convolutional neural network (CNN), the algorithm has performed effectively to detect depression in voice samples [9].

RESEARCH BACKGROUND

Various researchers have elaborately studied the phenomenon of depression. Discussion from few of those is as stated.

Karol Chalsta et al[10]suggested the method of deep convolutional neural networks for analyzing depression in speech. Network architectures were developed and tested for providing the best classification results. Audio spectrograms were obtained by training short voice samples on a model. The algorithm that was developed predicted an accuracy in the range of 77 % despite voice as the only feature being a barrier in analysis of depression [11].

Yasin Ozkanca et al[12]studied the problem of depression in people affected by Parkinson's disease. The paper shows similarity between the phenomenon of depression and Parkinson's disease. The emphasis here is on how feature analysis can be more accurate to detect depression, its complexity as well as accuracy.

Yanfei Wang et al [13] proposed a method of detecting depression using multiple instance learning. Here the authors aim to identify depression traits by getting the characteristic features of depressive patients by obtaining their facial features. The frames extracted from the features are trained in a model. The technique of featuring coupled with slicing and sampling. The method used is Sampling Slicing Long-term short-term memory multiple instance learning.

Zhenyu Liu et al[14] discussed on the approach of proposing a sensing system which can help clinicians in diagnosing a depression detection system. The analysis was done on a sample size of 300 people out of which 100 were depressive, 100 were healthy and other 100 were high risk people. the process of interview and questionnaires helped in feature selection. This eliminated the redundant data to reduce the complexity to train the model accurately. MFCC was used to extract the features from voice

MATERIALS AND METHODS

The model developed during the research is used to detect depression from a given voice sample. To develop the model, voice data has been taken from Kaggle which is placed in two folders. The first folder RAVDESS contains 1012 files which has 44 trials per actor [15].

The RAVDESS data has 24 actors amongst which 12 are female and other 12 are male who have spoken a sentence in north American accent. Emotions such as sad, happy, calm, angry are a part of the verbal data collected. Each data file has a unique 7-digit numerical identifier which associates it to an emotion [16][17].

The other data from TESS has a target of 200 words spoken by two actresses whose voice represent 7 different emotions however the voice samples from TESS are not depicted by a particular number signifying emotion. In order to associate each voice sample with an emotion a graphical user interface has been used.

A Graphical user interface is developed to enlarge the data and give a number to the voice sample which depicts an emotion. The GUI developed is as seen in figure 1. It collects the features of voice data from features folder which is placed at a particular location. Also, it collects data from another folder called TESS whose path needs to be specified. The data collected through this graphical user interface is now enhanced.

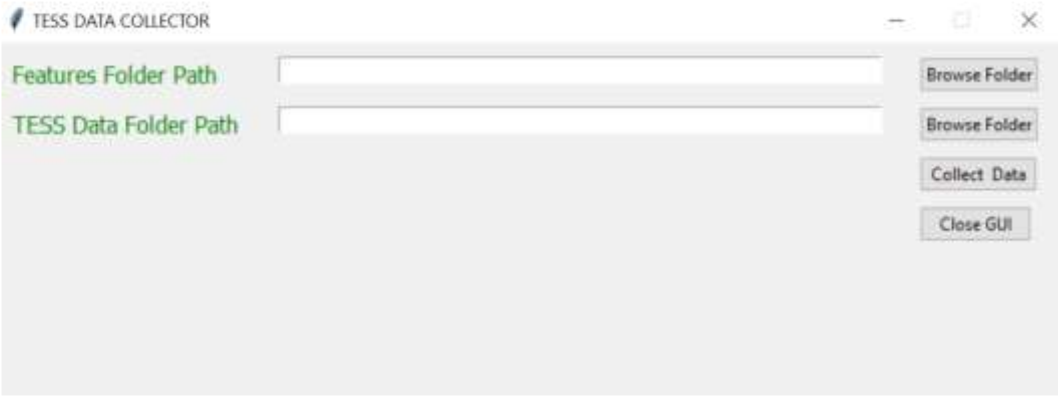


Figure 1: Graphical User Interface

The data obtained is then used to train the model using CNN. We can provide any voice sample to our trained model, whose features are extracted by applying the technique of Mel frequency Cepstral coefficients.

Pros: There have been various models that predict the emotion from voice data, but this model can depict depression based on emotions of a person, as the model has been trained on large data set of emotions. The accuracy score of the model is nearly 80% which means that the problem of over fitting of data is also avoided in the model as it predicts a nearby accurate result.

Cons: Since data available on depression is limited the model could have been trained better if a larger amount of data was there, as it would have helped in depicting the mood state of the person accurately.

RESULTS AND DISCUSSION

The audio data which is available to train the model has been characterized against 8 emotion parameters as shown in table 1.

Table1: Numerical representation of emotions

Number	Emotion
1	Neutral
2	Calm
3	Happy
4	Sad
5	Angry
6	Fearful
7	Disgust
8	Surprised

These audio files when collected from two sources and enhanced using GUI which is shown in figure 1 allows us to train the model as the data set is now prepared.

The audio file which is the test data that is fed to the model, its features are pre -processed using the MFCC and we obtain a Mel Spectrogram for the voice sample as seen in figure 2.

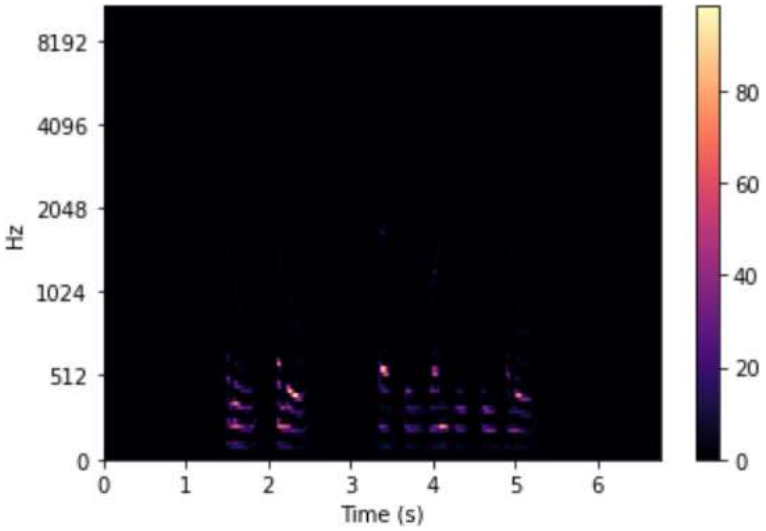


Figure 2: Mel spectrogram for audio sample

MFCC converts the frequency channel into power spectrum. This is done to enhance the vocal features and to minimize any loss in the voice sample or eliminate outside voices which might be a voice of the recorded sample. The power spectrum is then obtained as shown in figure 3.

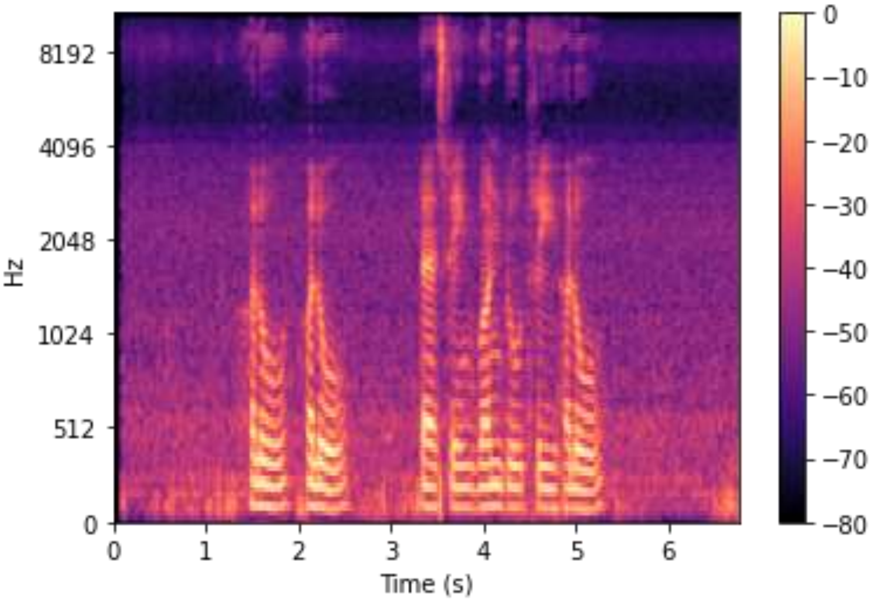
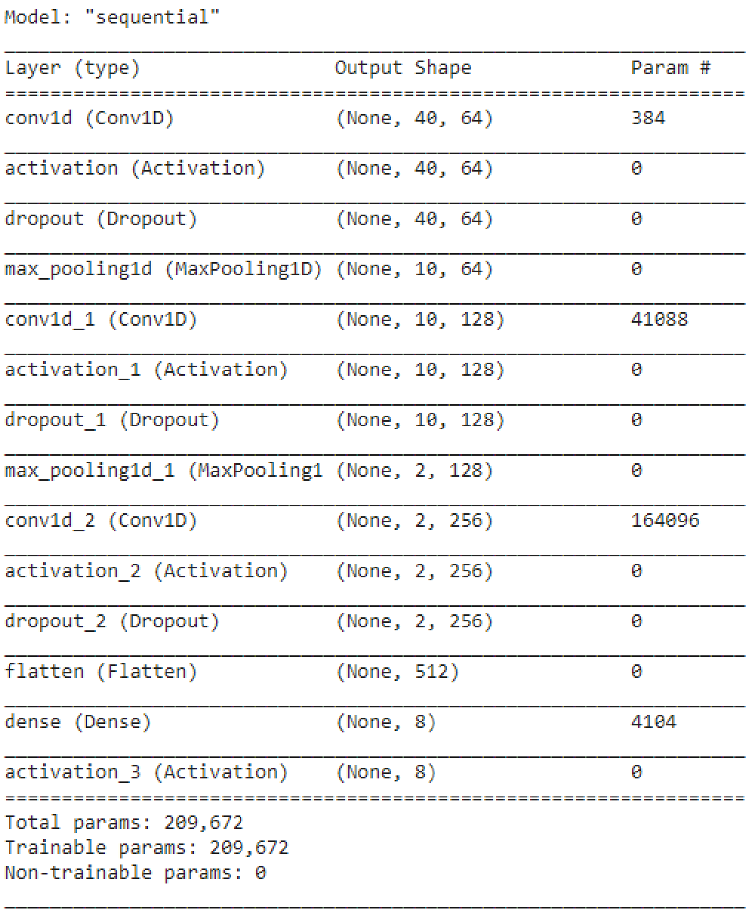


Figure 3: Power Spectrogram of voice sample.

All the audio data that we need to train the model is stored and MFCC's are collected for the entire sample range. Since the model has to start its prediction from 0, conversion of emotion features which was from 1 to 8 is done for a numbering of 0 to 7, each numbering representing the same emotion as earlier.



```
Model: "sequential"
_________________________________________________________________
Layer (type)                 Output Shape              Param #   
=================================================================
conv1d (Conv1D)              (None, 40, 64)            384       
_________________________________________________________________
activation (Activation)      (None, 40, 64)            0         
_________________________________________________________________
dropout (Dropout)            (None, 40, 64)            0         
_________________________________________________________________
max_pooling1d (MaxPooling1D) (None, 10, 64)            0         
_________________________________________________________________
conv1d_1 (Conv1D)            (None, 10, 128)           41088     
_________________________________________________________________
activation_1 (Activation)    (None, 10, 128)           0         
_________________________________________________________________
dropout_1 (Dropout)          (None, 10, 128)           0         
_________________________________________________________________
max_pooling1d_1 (MaxPooling1 (None, 2, 128)            0         
_________________________________________________________________
conv1d_2 (Conv1D)            (None, 2, 256)            164096    
_________________________________________________________________
activation_2 (Activation)    (None, 2, 256)            0         
_________________________________________________________________
dropout_2 (Dropout)          (None, 2, 256)            0         
_________________________________________________________________
flatten (Flatten)            (None, 512)               0         
_________________________________________________________________
dense (Dense)                (None, 8)                 4104      
_________________________________________________________________
activation_3 (Activation)    (None, 8)                 0         
=================================================================
Total params: 209,672
Trainable params: 209,672
Non-trainable params: 0
_________________________________________________________________
```

Figure 4: Sequential model summary

The figure 4 represents the sequential model summary of the model that has been obtained after training using Convolutional Neural networks. The CNN technique does layer on the model and defines the size of each layer as shown in above figure.

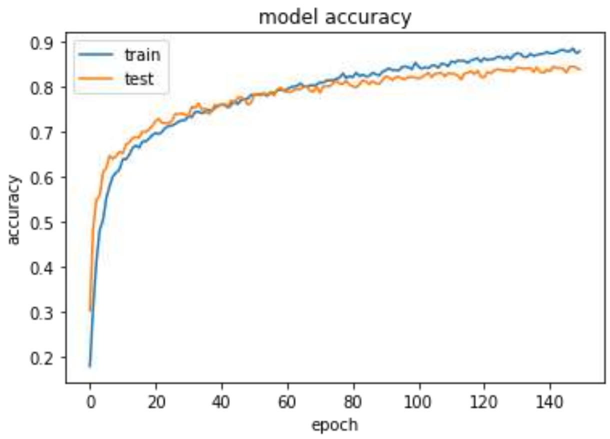


Figure 5: Model accuracy

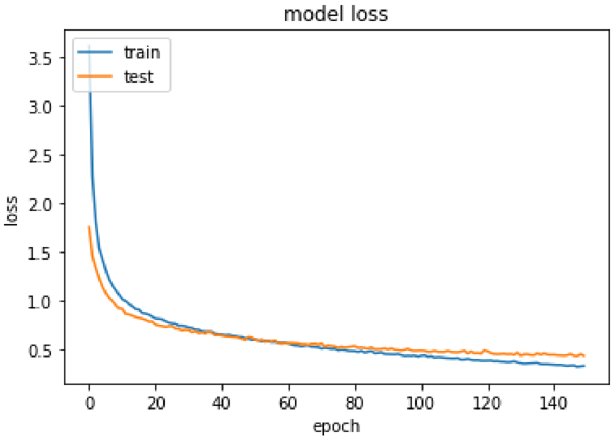
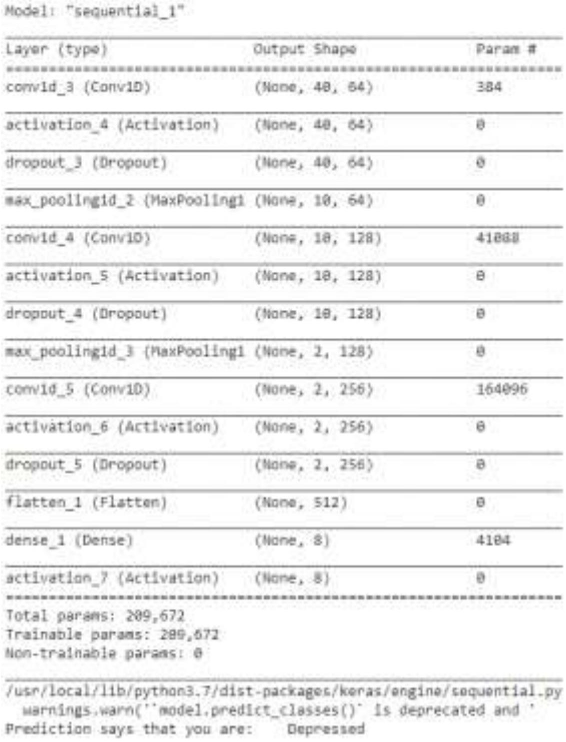


Figure 6: Model loss

Figure 5 and 6 depict that as epochs i.e., generations are generated while training the model it is observed that accuracy gets increased during each successive epoch and the model loss is minimized. This is a pro as compared to other models where the problem of accuracy is seen, wherein either the model works for an overfitted data or the accuracy is too low.



```
Model: "sequential_1"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
conv1d_3 (Conv1D)            (None, 40, 64)            384
_________________________________________________________________
activation_4 (Activation)    (None, 40, 64)            0
_________________________________________________________________
dropout_3 (Dropout)          (None, 40, 64)            0
_________________________________________________________________
max_pooling1d_2 (MaxPooling1 (None, 10, 64)            0
_________________________________________________________________
conv1d_4 (Conv1D)            (None, 10, 128)           41088
_________________________________________________________________
activation_5 (Activation)    (None, 10, 128)           0
_________________________________________________________________
dropout_4 (Dropout)          (None, 10, 128)           0
_________________________________________________________________
max_pooling1d_3 (MaxPooling1 (None, 2, 128)            0
_________________________________________________________________
conv1d_5 (Conv1D)            (None, 2, 256)            164096
_________________________________________________________________
activation_6 (Activation)    (None, 2, 256)            0
_________________________________________________________________
dropout_5 (Dropout)          (None, 2, 256)            0
_________________________________________________________________
flatten_1 (Flatten)          (None, 512)               0
_________________________________________________________________
dense_1 (Dense)              (None, 8)                 4104
_________________________________________________________________
activation_7 (Activation)    (None, 8)                 0
=================================================================
Total params: 209,672
Trainable params: 209,672
Non-trainable params: 0
_________________________________________________________________
/usr/local/lib/python3.7/dist-packages/keras/engine/sequential.py
  warnings.warn('`model.predict_classes()` is deprecated and '
Prediction says that you are:    Depressed
```

Figure 7: Updated sequential model with prediction

Figure 7 depicts the updated sequential model which depicts the total number of trainable and non-trainable parameters, with different layers of CNN and when a sample test file is fed to this model, the model shall accurately predict whether the person is in a depressed state or is not depressed.

CONCLUSION

In this paper a depression detection model is generated using Mel frequency cepstral coefficients and Convolutional Neural Networks approach which detects the severity of depression from a voice sample. The audio sample is preprocessed and normalized using MFCC and the coefficients obtained are then fed to the model developed by CNN. The problem of overfitting has been avoided in the model and the accuracy is enhanced while keeping the loss as minimal as possible.

In future we can enhance up the same model with a web application that can automatically diagnose depressive patients.

REFERENCES

1. World Health Organization (2017) “WHO global health days - Staying positive and preventing depression as you get older.” Retrieved from https://www.who.int/campaigns/worldhealthday/2017/handoutsdepression/older-age/en/

2. Snell-Rood, Claire, Richard Merkel, and Nancy Schoenberg. "Negotiating the interpretation of depression shared among kin." Medical anthropology 37, no. 7 : 538-552,2018.

3. Fjermestad-Noll, Jane, Elsa Ronningstam, Bo Bach, Bent Rosenbaum, and Erik Simonsen. "Characterological depression in patients with narcissistic personality disorder." Nordic journal of psychiatry 73, no. 8 : 539-545, 2019.

4. WHO. The global burden of disease: 2004 update. Geneva: WHO; 2004 [accessed 2019 Apr 9]. Available from : https://www.who.int/healthinfo/global_burden_disease/2004_report_update/en/.

5. Rihmer, Z. "Can better recognition and treatment of depression reduce suicide rates? A brief review." European Psychiatry 16, no. 7: 406-409, 2001.

6. World Health Organization. (2018) “Agenda Item 12.4 Digital Health.” In Proceedings of Seventy-First World Health Assembly, 21–26 May 2018 in Geneva,Switzerland(pp.3)http://apps.who.int/gb/ebwha/pdf_files/WHA71/A71_R7-en.pdf

7. Quatieri, Thomas F., and Nicolas Malyska. "Vocal-source biomarkers for depression: A link to psychomotor activity." In Thirteenth Annual Conference of the International Speech Communication Association, 2012.

8. Tiwari, Vibha. "MFCC and its applications in speaker recognition." International journal on emerging technologies 1, no. 1: 19-22, 2010.

9. Yang, Le, Hichem Sahli, Xiaohan Xia, Ercheng Pei, Meshia Cédric Oveneke, and Dongmei Jiang. "Hybrid depression classification and estimation from audio video and text information." In Proceedings of the 7th annual workshop on audio/visual emotion challenge, pp. 45-51,2017.

10. Chlasta K, Wołk K, Krejtz I. Automated speech-based screening of depression using deep convolutional neural networksProcediaComputerScience,164:618-28, Jan 1 2019.

11. Afshan, Amber, Jinxi Guo, Soo Jin Park, Vijay Ravi, Jonathan Flint, and Abeer Alwan. "Effectiveness of voice quality features in detecting depression." Interspeech,2018.

12. Ozkanca, Yasin, Miraç Göksu Öztürk, Merve Nur Ekmekci, David C. Atkins, Cenk Demiroglu, and Reza Hosseini Ghomi. "Depression screening from voice samples of patients affected by parkinson’s disease." Digital biomarkers 3, no. 2 : 72-82,2019.

13. Wang, Yanfei, Jie Ma, Bibo Hao, Pengwei Hu, Xiaoqian Wang, Jing Mei, and Shaochun Li. "Automatic Depression Detection via Facial Expressions Using Multiple Instance Learning." In 2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI), pp. 1933-1936. IEEE,2020.

14. Liu, Zhenyu, Bin Hu, Lihua Yan, Tianyang Wang, Fei Liu, Xiaoyu Li, and Huanyu Kang. "Detection of depression in speech." In 2015 international conference on affective computing and intelligent interaction (ACII), pp. 743-747. IEEE,2015.

15. Livingstone, Steven R., and Frank A. Russo. "The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English." PloS one 13, no. 5: e0196391,2018.

16. Ouyang, Xi, Shigenori Kawaai, Ester Gue Hua Goh, Shengmei Shen, Wan Ding, Huaiping Ming, and Dong-Yan Huang. "Audio-visual emotion recognition using deep transfer learning and multiple temporal models." In Proceedings of the 19th ACM International Conference on Multimodal Interaction, pp. 577-582,2017.

17. Atalay, Tolga, Deger Ayata, and Yusuf Yaslan. "Comparison of feature selection methods in voice based emotion recognition systems." In 2018 26th Signal Processing and Communications Applications Conference (SIU), pp. 1-4. IEEE, 2018.