ORIGINAL RESEARCH ARTICLE

Deep learning model for identification of customers satisfaction in business

Sangeetha Ganesan

Department of Artificial Intelligence and Data Science, R.M.K College of Engineering and Technology, Tiruvallur District, Tamil Nadu 601206, India; gsangeethakarthik@gmail.com

ABSTRACT

Communication is the key to expressing one's feelings and ideas evidently. Recognition of the emotional state of a speaker is a significant step in making the human machine communication more natural and approachable. The knowledge behind creating this work was to make a deep learning model that might identify the emotions from their speech. This can be used by multiple industries to offer different services like marketing company signifying you to buy products based on your emotions, automotive industry can detect the customer's emotions and adjust the speed of self-directed cars as required to avoid any collisions, etc. The proposed system has involved classifying the emotion into angry, calm, fear, happy and sad categories from the audio signals using classifier algorithms like MultiLayer Perceptron and Convolutional Neural Network. The customer satisfaction is very important to enhance the business. By using this model the customer can identify the customer satisfaction. The suggested methods open the door for a real-time prototype for customer speech emotion recognition with open-source features to boost business profitability.

Keywords: customer satisfaction; Convolutional Neural Network; speech emotion recognition; classification report MultiLayer Perceptron

ARTICLE INFO

Received: 30 June 2023
Accepted: 15 August 2023
Available online: 13 October 2023

1. Introduction

Morse defined that "Satisfaction means the stage of achievement of the needs of a person, desires and wants[1]. Satisfaction depends upon what a person needs from the world, and what she obtains." Many organization theorists and experts recommend corporations to concentrate on their customers' requirements and satisfaction. It is a general to tactical organization, the hunt of "excellence", the Marketing model, total quality management board, service quality theorists market orientation, and association with marketing policies. Conversely, in spite of the availability of many techniques and systems for monitoring and measuring customer satisfaction and utilizing the decision making, there are most important achievement issues facing a customer satisfaction policy that have been absolutely disregarded.

A research topic called "Speech Emotion Detection" seeks to extract the sentiment from the speech emotion. Numerous examinations state that development in emotion detection will make a lot of systems easier and hence make the world a better place to live. Humans utilize most parts of their body and voice to communicate effectively. Hand motions, body language, and the tone and temperament are all collectively used to express one's feelings. Though the verbal part of the communication varies by languages practiced across the globe, the non-verbal part of communication is the expression of feeling which is mostly common among all people. Nearly there are many research areas that are getting advantages from systematizing the emotion detection technique including psychology, psychiatry, and neuroscience. A potential downside occurs as few people are classified as introverts and hesitate to communicate. The objective of this work is to develop a speech sentiment recognition system that improves man-machine interface and to identify the emotional state of the customer to improve the business profit. This system uses a convolutional network that inputs the audio signals and outputs the emotional state of the customer.

2. Background

Customer satisfaction is declared as the emotional state subsequent when the feeling surrounding disconfirmed anticipation is attached with the previous feelings of customers concerning the expenditure knowledge[2]. Customers frequently build a choice to purchase or repurchase after estimating whether their knowledge with the service or product has been pleasurable or satisfactory[3,4]. The customer satisfaction form depended on the anticipation disconfirmation hypothesis sponsors that customers are satisfied when definite hard presentation outperforms or confirms ancient times. Disconfirmation happens when there are dissimilarities between outcomes and expectations. Negative disconfirmation happens when service or product presentation is inferior than anticipated while positive disconfirmation happens when service or product or presentation is enhanced than anticipated. Positive confirmation and disconfirmation or outcomes in customer satisfaction while negative disconfirmation guides to customer displeasure.

Khare and Bajaj[5] proposed an emotion recognition model that has attracted wide benefits in affective computing, medical, brain-computer interface, and other relevant fields. Emotion recognition from Electroencephalogram (EEG) signals has facilitated to impaired people to be intact with the real world. The Eigenvector Centrality Method (EVCM) is applied for dominant channel selection. For the adaptive disintegration of non-stationary EEG signals, the optimal combination of modes and penalty factors is chosen. A fantastic learning machine classifier for a four-emotion classification achieves an overall accuracy of 71.24%, he suggested approach performs 4% and 2% better on the same dataset than the state-of-the-art and conventional variational mode decomposition, respectively[5]. Wani et al.[6] proposed a system that goals the customer's survival of different emotions by removing and classifying the well-known feature from a pre-processed speech signal. The method machines and humans identify and associate touching features of speech signals are relatively distinct qualitatively and quantitatively, which present huge complexities specifically in speech feeling acknowledgment and human computer interface[6].

A new outline for speech feeling acknowledgment was proposed by Mustaqeem et al., using a key progression part selection that was based on group-based Redial Based Function Network (RBFN) correspondence measurements. The Short-Time Fourier Transform (STFT) algorithm is used to transform the chosen sequence into a spectrogram, which is then passed into the Convolutional Neural Network (CNN) model to extract the salient and discriminative features from the speech spectrogram. The experiments show that the proposed speech feeling reorganisation model is robust and effective when compared to state-of-the-art speech feeling reorganisation techniques, achieving up to 72.25% accuracy over Interactive Emotional Dyadic Motion Capture (IEMOCAP)[7]. A Parallelized Convolutional Recurrent Neural Network (PCRN) for reorganising speech feelings was created by Jiang et al. and Nwe et al. Emotion classification is done using a SoftMax classifier. To aid in the understanding of subtle changes in emotion, the PCRN structure processes two dissimilar kinds of attributes simultaneously[8,9].

A new model developed by Wang et al. predicts emotions based on Mel-Frequency Cepstral Coefficients (MFCC) characteristics and Mel spectrograms made from unprocessed signals. Each utterance goes through a pre-processing step that produces two Mel-spectrograms and MFCC attributes with different instance regular resolutions. In contrast to traditional Long Short-Term Memory (LSTMs), which process the MFCC features, the Dual-Sequence LSTM (DS-LSTM) architecture is a novel LSTM that simultaneously processes two Mel-spectrograms. This proposed model outperforms state-of-the-art unimodal models by 6%, multimodal models that combine both textual and audio signals by 73.3%, and state-of-the-art unimodal models by 72.7% and 73.3%, respectively[10,11]. To identify various human emotions from speech signals, Chattopadhyay et al.[12] created CEOAS (Clustering-based Equilibrium Optimizer (EO)) hybrid wrapper feature selection algorithm and Atom Search Optimisation (ASO). From the audio signals, they have taken Linear Prediction Coding (LPC) and Linear Predictive Cepstral Coefficient (LPCC). The South Asian Association for Regional Cooperation (SAARC) Audio-visual Exchange, the Berlin Emotional Speech Database, the Ryerson Audio Visual Database of Emotional Speech and Song (RAVDESS), and IEMOCAP were used as four benchmark datasets that are accepted in the industry. We achieved impressive recognition accuracies of 74.25%, outperforming many state-of-the-art algorithms[12,13].

The Tensor Factorised Neural Networks (TFNN) and Attention Gated Tensor Factorised Neural Network (AG-TFNN) speech emotion recognition were proposed by Pandey et al. Because Mel spectrograms are 2D tensors by nature, TFNN and AG-TFNN are a better option than baselines like Convolutional Neural Network and Long Short-Term Memory because they have fewer parameters to learn and a simpler architecture. Research work performed on emotional speech datasets IEMOCAP and Emotional Speech database proves that TFNN and AG-TFNN are better than the state-of-the-art given by CNN + LSTM combination with fewer numbers of parameters[14].

In order to address the issue of speech feeling detection, Pande et al.[15] proposed a new structural design that can benefit from the advantages of various networks while overcoming the drawbacks of using just one. This structural design has been tested on the well-known IEMOCAP database and Berlin Database of Emotional Speech corpus. In both the speaker-dependent experiment and the speaker-independent experiment, they outperformed previous state-of-the-art methods in the spontaneous emotional speech of the IEMOCAP data by 74.96% unweighted accuracy. This is an impressive accomplishment. They employed the 3-D Log-Mel spectra values from the raw signals to convert emotional dialogues into spectrograms. Using the aforementioned datasets, they conducted an experiment to assess the robustness and generalizability, and the results showed a superior recognition accuracy of 63.84%[15].

3. Proposed work

The goal of this work was to develop a deep learning model that could recognize many human emotions from a customer's speech, such as anger, sadness, husky voice, happiness, disgust, surprise, neutrality, and fear. The first step is to convert the speech to text mode using speech emotion Application Programming Interface (API) to know what the audio is all about. This is done by using recognizers. The usage of neural networks provides the advantage of classifying many different types of emotion in variable length of audio signal in a real time environment. After converting audio to text format masking and cleaning of audio files are done to remove unwanted background noise, time stretching audio and pitch of audio files. MFCC, Mel, chroma, contrast, tonnetz has been used as the feature for classifying the speech data into various emotion categories employing artificial neural networks. Then the data are split into training and test samples. This method used MultiLayer Perceptron and Convolutional Neural Network as the classifier algorithm. This method directs to set up a good balance between performance accuracy of the real-time processes and computational volume. Then the live test to record the audio and print the speech text format is done. The process of the proposed system (**Figure 1**) is as follows,

1) Initially the audio files are preprocessed, i.e., masking and cleaning of audio files to remove unwanted background noises from the audio files.
2) Next the audio files are converted to text format using speech recognizer.
3) Then the necessary features like chroma, MFCC, Mel, tonnetz are taken out from the audio source.
4) After that the data are pre-split into the train and test set.

5) The next step is to build and train a Deep Neural Network with our clean and prepared dataset. For this model, we decided to use classification algorithms MultiLayer Perceptron and 1D CNN as we have a time dimension aspect in our audio features.

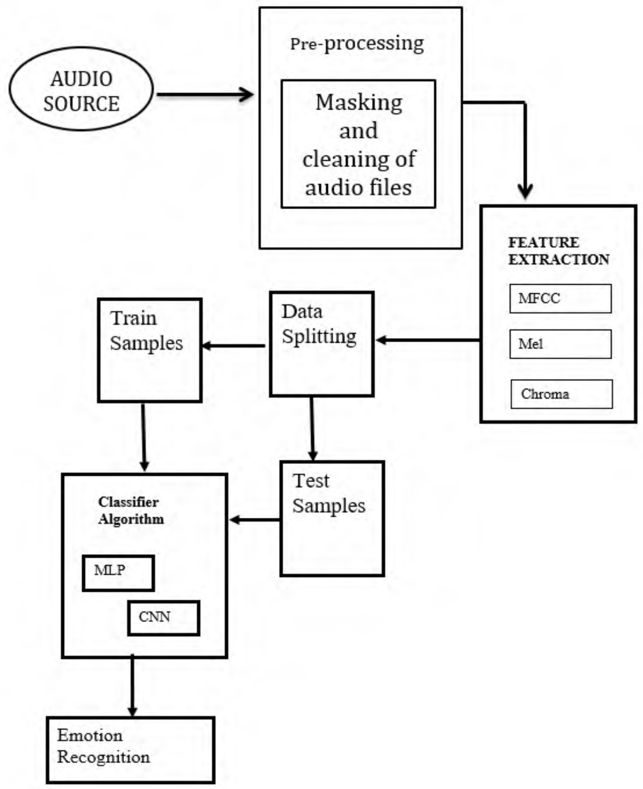


Figure 1. Process of the proposed system.

4. Experimental result and performance analysis

A processor of 500 MHz with 64-bit Windows 10 of 16GB Random Access Memory (RAM) System is taken for the implementation. Anaconda Navigator and Python IDE 3.7.4 (Jupyter Notebook) are used with the library of NumPy, Matplotlib, TensorFlow, ReLU, Keras, Scipy, sklearn, Librosa, Pandas.

Initially loaded the folders containing the speech from the Ryerson Audio-Visual Database of Emotional Speech and dataset which contains voices of 20 customers and there were totally 1140 audio files of both male and female customers.

The wave plots of audio with noisy background and the time stretching audio are given in **Figures 2** and **3** respectively.

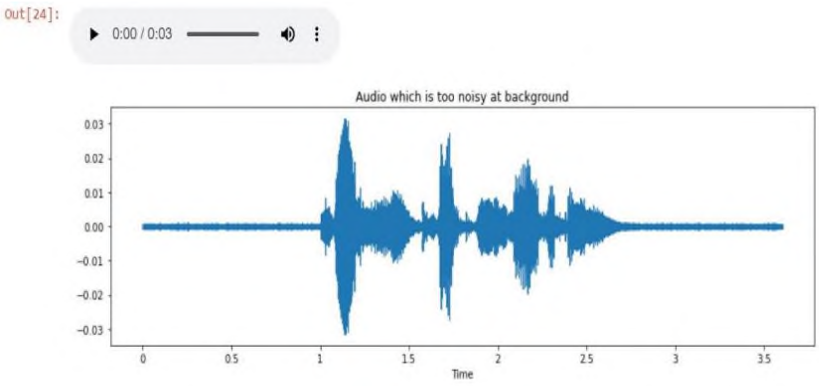
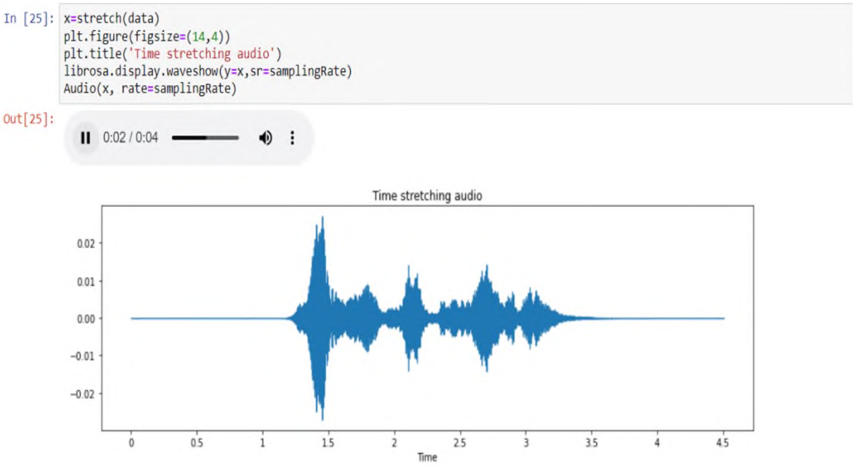


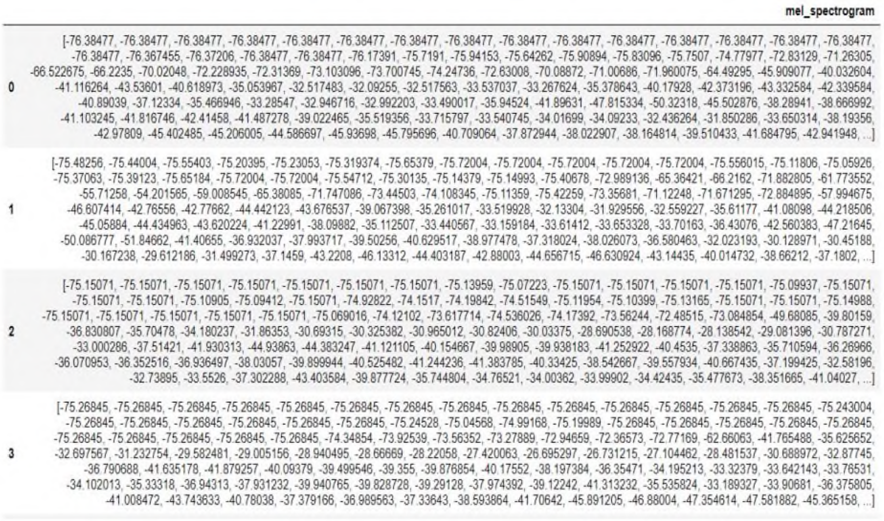
Figure 2. Wave plot of audio with noisy background.



```
In [25]: x=stretch(data)
         plt.figure(figsize=(14,4))
         plt.title('Time stretching audio')
         librosa.display.waveshow(y=x,sr=samplingRate)
         Audio(x, rate=samplingRate)

Out[25]:
```

Figure 3. Time stretching audio.

A Fast Fourier Transform is calculated on overlie windowed sections of the signal is known as spectrogram. This spectrogram depicts amplitude which is mapped on a Mel scale. The below **Figure 4** represents the extraction of features using Mel spectrogram.



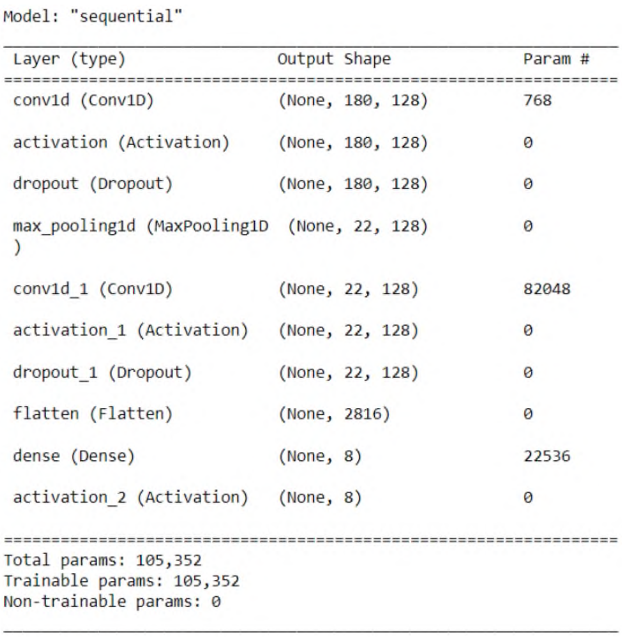
	mel_spectrogram
0	[-76.38477, -76.38477, -76.38477, -76.38477, -76.38477, -76.38477, -76.38477, -76.38477, -76.38477, -76.38477, -76.38477, -76.38477, -76.38477, -76.38477, -76.38477, -76.38477, -76.367455, -76.37206, -76.38477, -76.38477, -76.17391, -75.7191, -75.94153, -75.64262, -75.90894, -75.83096, -75.7507, -74.77977, -72.83129, -71.26305, -66.522675, -66.2235, -70.02048, -72.228935, -72.31369, -73.103096, -73.700745, -74.24736, -72.63008, -70.08872, -71.00686, -71.960075, -64.49295, -45.909077, -40.032604, -41.116264, -43.53601, -40.618973, -35.053967, -32.517483, -32.09255, -32.517563, -33.537037, -33.267624, -35.378643, -40.17928, -42.373196, -43.332584, -42.339584, -40.89039, -37.12334, -35.466946, -33.28547, -32.946716, -32.992203, -33.490017, -35.94524, -41.89631, -47.815334, -50.32318, -45.502876, -38.28941, -38.666992, -41.103245, -41.816746, -42.41458, -41.487278, -39.022465, -35.519356, -33.715797, -33.540745, -34.01699, -34.09233, -32.436264, -31.850286, -33.650314, -38.19356, -42.97809, -45.402485, -45.206005, -44.586697, -45.93698, -45.795696, -40.709064, -37.872944, -38.022907, -38.164814, -39.510433, -41.684795, -42.941948, ...]
1	[-75.48256, -75.44004, -75.55403, -75.20395, -75.23053, -75.319374, -75.65379, -75.72004, -75.72004, -75.72004, -75.72004, -75.72004, -75.556015, -75.11806, -75.05926, -75.37063, -75.39123, -75.65184, -75.72004, -75.72004, -75.54712, -75.30135, -75.14379, -75.14993, -75.40678, -72.989136, -65.36421, -66.2162, -71.882805, -61.773552, -55.71258, -54.201565, -59.008545, -65.38085, -71.747086, -73.44503, -74.108345, -75.11359, -75.42259, -73.35681, -71.12248, -71.671295, -72.884895, -57.994675, -46.607414, -42.76556, -42.77662, -44.442123, -43.676537, -39.067398, -35.261017, -33.519928, -32.13304, -31.929556, -32.559227, -35.61177, -41.08098, -44.218506, -45.05884, -44.434963, -43.620224, -41.22991, -38.09882, -35.112507, -33.440567, -33.159184, -33.61412, -33.653328, -33.70163, -36.43076, -42.560383, -47.21645, -50.086777, -51.84662, -41.40655, -36.932037, -37.993717, -39.50256, -40.629517, -38.977478, -37.318024, -38.026073, -36.580463, -32.023193, -30.128971, -30.45188, -30.167238, -29.612186, -31.499273, -37.1459, -43.2208, -46.13312, -44.403187, -42.88003, -44.656715, -46.630924, -43.14435, -40.014732, -38.66212, -37.1802, ...]
2	[-75.15071, -75.15071, -75.15071, -75.15071, -75.15071, -75.15071, -75.15071, -75.13959, -75.07223, -75.15071, -75.15071, -75.15071, -75.15071, -75.09937, -75.15071, -75.15071, -75.15071, -75.10905, -75.09412, -75.15071, -74.92822, -74.1517, -74.19842, -74.51549, -75.11954, -75.10399, -75.13165, -75.15071, -75.15071, -75.14988, -75.15071, -75.15071, -75.15071, -75.15071, -75.15071, -75.069016, -74.12102, -73.617714, -74.536026, -74.17392, -73.56244, -72.48515, -73.084854, -49.68085, -39.80159, -36.830807, -35.70478, -34.180237, -31.86353, -30.69315, -30.325382, -30.965012, -30.82406, -30.03375, -28.690538, -28.168774, -28.138542, -29.081396, -30.787271, -33.000286, -37.51421, -41.930313, -44.93863, -44.383247, -41.121105, -40.154667, -39.98905, -39.938183, -41.252922, -40.4535, -37.338863, -35.710594, -36.26966, -36.070953, -36.352516, -36.936497, -38.03057, -39.899944, -40.525482, -41.244236, -41.383785, -40.33425, -38.542667, -39.557934, -40.667435, -37.199425, -32.58196, -32.73895, -33.5526, -37.302288, -43.403584, -39.877724, -35.744804, -34.76521, -34.00362, -33.99902, -34.42435, -35.477673, -38.351665, -41.04027, ...]
3	[-75.26845, -75.26845, -75.26845, -75.26845, -75.26845, -75.26845, -75.26845, -75.26845, -75.26845, -75.26845, -75.26845, -75.26845, -75.26845, -75.26845, -75.243004, -75.26845, -75.26845, -75.26845, -75.26845, -75.26845, -75.26845, -75.24528, -75.04568, -74.99168, -75.19989, -75.26845, -75.26845, -75.26845, -75.26845, -75.26845, -75.26845, -75.26845, -75.26845, -75.26845, -75.26845, -74.34854, -73.92539, -73.56352, -73.27889, -72.94659, -72.36573, -72.77169, -62.66063, -41.765488, -35.625652, -32.697567, -31.232754, -29.582481, -29.005156, -28.940495, -28.66669, -28.22058, -27.420063, -26.695297, -26.731215, -27.104462, -28.481537, -30.688972, -32.87745, -36.790688, -41.635178, -41.879257, -40.09379, -39.499546, -39.355, -39.876854, -40.17552, -38.197384, -36.35471, -34.195213, -33.32379, -33.642143, -33.76531, -34.102013, -35.33318, -36.94313, -37.931232, -39.940765, -39.828728, -39.29128, -37.974392, -39.12242, -41.313232, -35.535824, -33.189327, -33.90681, -36.375805, -41.008472, -43.743633, -40.78038, -37.379166, -36.989563, -37.33643, -38.593864, -41.70642, -45.891205, -46.88004, -47.354614, -47.581882, -45.365158, ...]

Figure 4. Feature extraction using Mel spectrogram.

Our CNN model was trained with 32 epochs and 100 batch size since this setup was more suitable with the sample amount of dataset we used. Group size is significant a hyper parameter to alter in deep learning. Conversely, it is known that too huge of a group size will lead to the worst generalization. Various epochs were experimented to obtain the best performance results. The next step is to build the model. Keras (Neural Networks API) provides a way to summarize a model. The outline can be made by invoking the summary function on the system that gets a sequence that in turn can be displayed. The summary is textual and includes information about:

- The layers and the order in the model.
- The output shape of each layer.
- The number of parameters (weights) in each layer.
- The entire number of arguments in the model (weights).

The model summary is given in **Figure 5**.



```
Model: "sequential"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 conv1d (Conv1D)             (None, 180, 128)          768

 activation (Activation)     (None, 180, 128)          0

 dropout (Dropout)           (None, 180, 128)          0

 max_pooling1d (MaxPooling1D  (None, 22, 128)          0
 )

 conv1d_1 (Conv1D)           (None, 22, 128)           82048

 activation_1 (Activation)   (None, 22, 128)           0

 dropout_1 (Dropout)         (None, 22, 128)           0

 flatten (Flatten)           (None, 2816)              0

 dense (Dense)               (None, 8)                 22536

 activation_2 (Activation)   (None, 8)                 0

=================================================================
Total params: 105,352
Trainable params: 105,352
Non-trainable params: 0
_________________________________________________________________
```

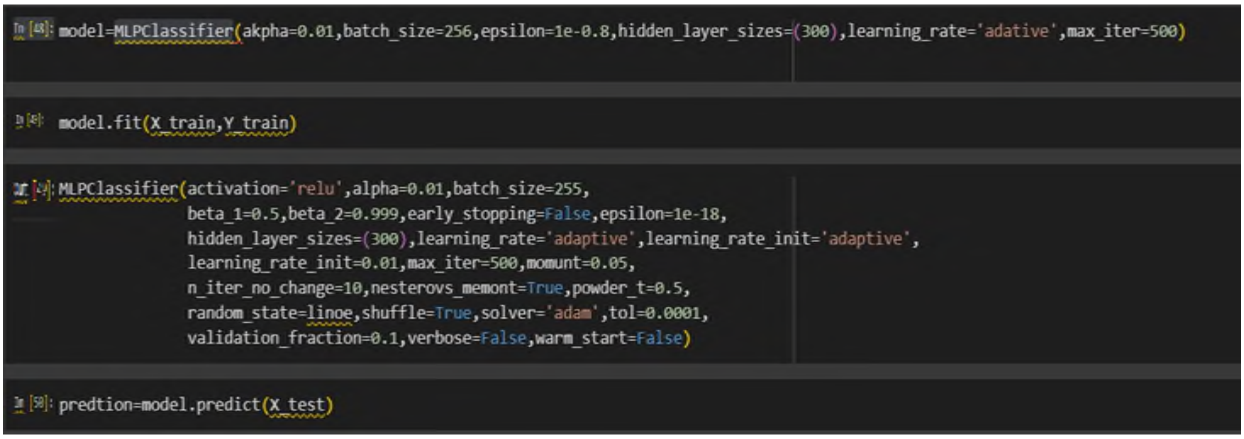
Figure 5. Model summary.

Building and tuning a model is a very time-consuming process. Always start basic and avoid adding too many layers merely to make something complex. After layer testing, the MultiLayer Perceptron and Convolutional Neural Network models that provided the highest validation accuracy against test data were 73% and 59%, respectively. As the graph states that both "training and testing" mistakes decrease as the number of epochs to the training model increases. The performance of the two classifier algorithms was measured and mentioned below.

4.1. MultiLayer Perceptron

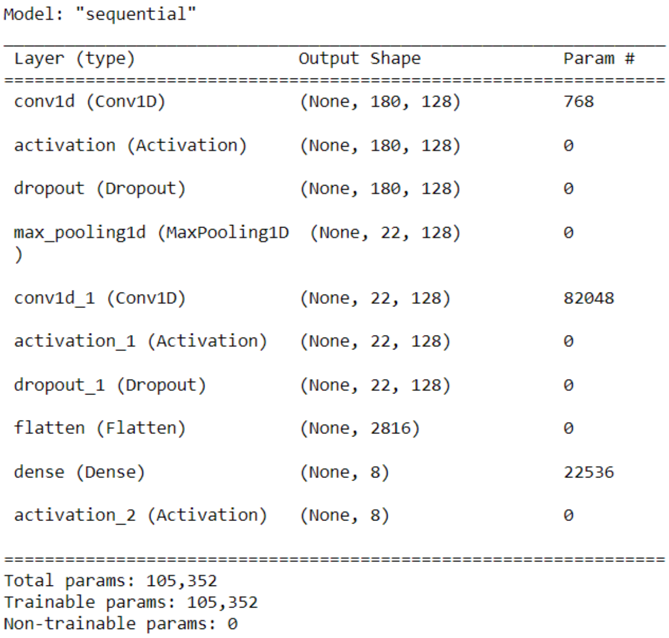
The Multi Layer Perceptron (MLP) Classifier function is used in predicting the emotion from the training and test set using the parameters alpha, batch_size and epsilon. Accuracy of predicting the emotion is printed using accuracy_score. It gives an accuracy of 75.9%.

The model building using MLP classifier and summary is given in **Figures 6** and **7**.



```
In [48]: model=MLPClassifier(akpha=0.01,batch_size=256,epsilon=1e-0.8,hidden_layer_sizes=(300),learning_rate='adative',max_iter=500)

In [49]: model.fit(X_train,Y_train)

Out[49]: MLPClassifier(activation='relu',alpha=0.01,batch_size=255,
                       beta_1=0.5,beta_2=0.999,early_stopping=False,epsilon=1e-18,
                       hidden_layer_sizes=(300),learning_rate='adaptive',learning_rate_init='adaptive',
                       learning_rate_init=0.01,max_iter=500,momunt=0.05,
                       n_iter_no_change=10,nesterovs_memont=True,powder_t=0.5,
                       random_state=linoe,shuffle=True,solver='adam',tol=0.0001,
                       validation_fraction=0.1,verbose=False,warm_start=False)

In [50]: predtion=model.predict(X_test)
```

Figure 6. Model building using MLP classifier.



```
Model: "sequential"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 conv1d (Conv1D)             (None, 180, 128)          768

 activation (Activation)     (None, 180, 128)          0

 dropout (Dropout)           (None, 180, 128)          0

 max_pooling1d (MaxPooling1D  (None, 22, 128)          0
 )

 conv1d_1 (Conv1D)           (None, 22, 128)           82048

 activation_1 (Activation)   (None, 22, 128)           0

 dropout_1 (Dropout)         (None, 22, 128)           0

 flatten (Flatten)           (None, 2816)              0

 dense (Dense)               (None, 8)                 22536

 activation_2 (Activation)   (None, 8)                 0

=================================================================
Total params: 105,352
Trainable params: 105,352
Non-trainable params: 0
```

Figure 7. Model summary.

4.2. Convolutional Neural Network

cnn_results=CNN_model.fit (x_traincnn, y_train_lb, batch_size = 64, epochs = 25, verbose = 1, validation_data = (x_testcnn, y_test_lb)) is used in predicting the emotion from the audio files using CNN. Here, the parameter epoch is the number of passes of the entire training data set and Batch_size is the number of training utilized in one iteration, usually 32 or 64 and must be a power of 2. It gives an accuracy of 53%.

The count of emotions is shown in **Figure 8**.

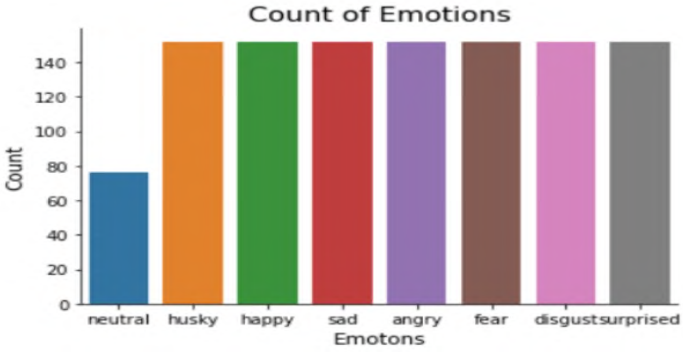
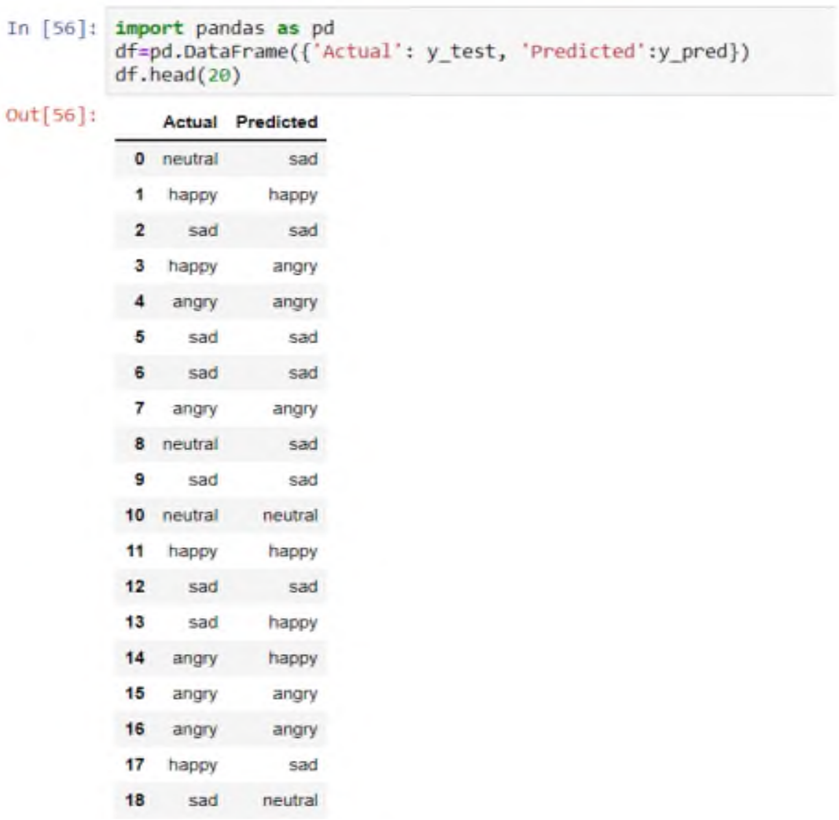


Figure 8. Count of emotions.

The actual and predicted emotion, and classification report are shown in **Figures 9** and **10**.

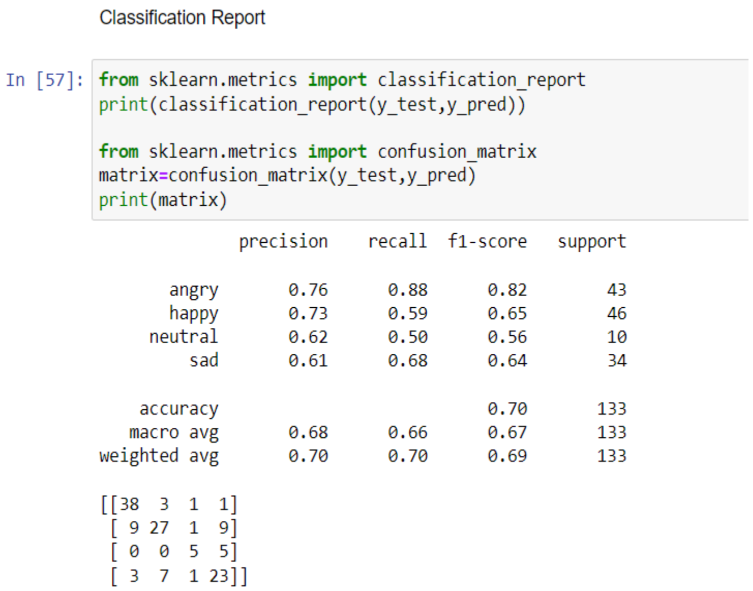


```
In [56]: import pandas as pd
         df=pd.DataFrame({'Actual': y_test, 'Predicted':y_pred})
         df.head(20)
```

Out[56]:

	Actual	Predicted
0	neutral	sad
1	happy	happy
2	sad	sad
3	happy	angry
4	angry	angry
5	sad	sad
6	sad	sad
7	angry	angry
8	neutral	sad
9	sad	sad
10	neutral	neutral
11	happy	happy
12	sad	sad
13	sad	happy
14	angry	happy
15	angry	angry
16	angry	angry
17	happy	sad
18	sad	neutral

Figure 9. Actual and predicted emotion.



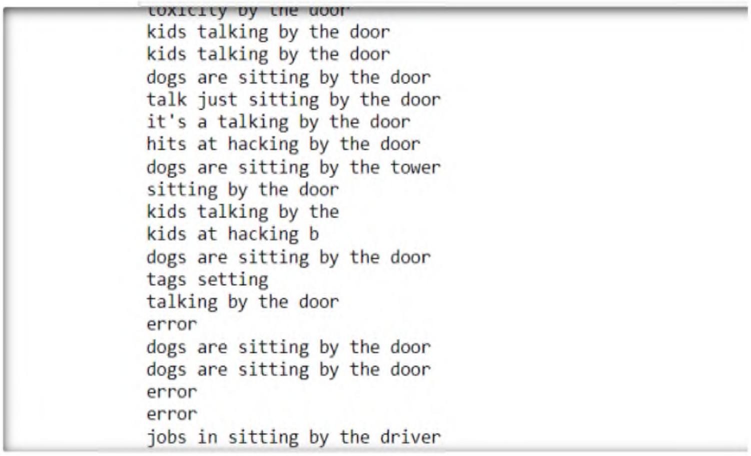
Classification Report

```
In [57]: from sklearn.metrics import classification_report
         print(classification_report(y_test,y_pred))

         from sklearn.metrics import confusion_matrix
         matrix=confusion_matrix(y_test,y_pred)
         print(matrix)
```

```
              precision    recall  f1-score   support

       angry       0.76      0.88      0.82        43
       happy       0.73      0.59      0.65        46
     neutral       0.62      0.50      0.56        10
         sad       0.61      0.68      0.64        34

    accuracy                           0.70       133
   macro avg       0.68      0.66      0.67       133
weighted avg       0.70      0.70      0.69       133

[[38  3  1  1]
 [ 9 27  1  9]
 [ 0  0  5  5]
 [ 3  7  1 23]]
```

Figure 10. Classification report.

The audio to text conversion is given in **Figure 11**. The sample output for husky voice is given in **Figure 12** and emotion detection is shown in **Figure 13**.



```
toxicity by the door
kids talking by the door
kids talking by the door
dogs are sitting by the door
talk just sitting by the door
it's a talking by the door
hits at hacking by the door
dogs are sitting by the tower
sitting by the door
kids talking by the
kids at hacking b
dogs are sitting by the door
tags setting
talking by the door
error
dogs are sitting by the door
dogs are sitting by the door
error
error
jobs in sitting by the driver
```

Figure 11. Audio to text conversion.

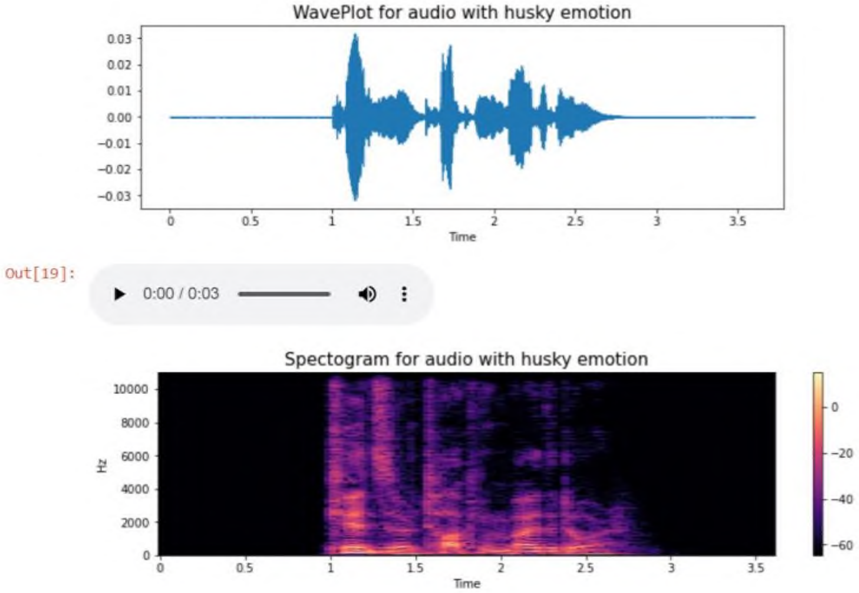


Figure 12. Output for husky voice.

```
Out[57]: array(['neutral'], dtype='<U7')
```

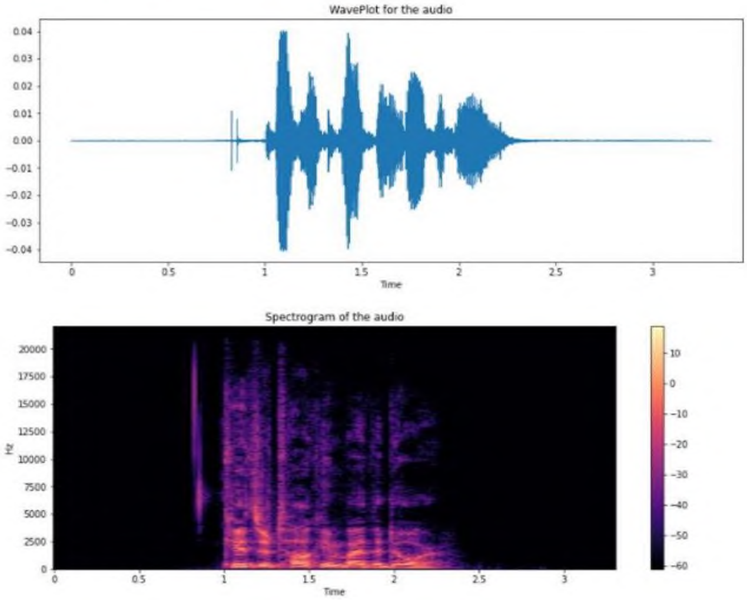


Figure 13. Emotion detection of a single audio file.

The above **Figure 13** shows the output of emotion from the particular audio file. After this we are also performing live audio recording and translating the live audio to text format. Precision is the relation of appropriately forecasted positive examinations to the whole forecasted positive examinations. Recall is the relation of appropriately forecasted positive examinations to all the observations in actual class. Precision is the majority instinctive performance assessment and it is basically a ratio of suitably forecasted examinations to the total examinations. CNN converges quicker than MLP but in terms of epoch CNN model gets extra time compared to MLP model as the number of parameters is more in CNN model. The accuracy for CNN model is 53% and for MLP is 75.9%. By comparing the accuracies of both, MLP has given better performance as shown in **Figures 14** and **15**.

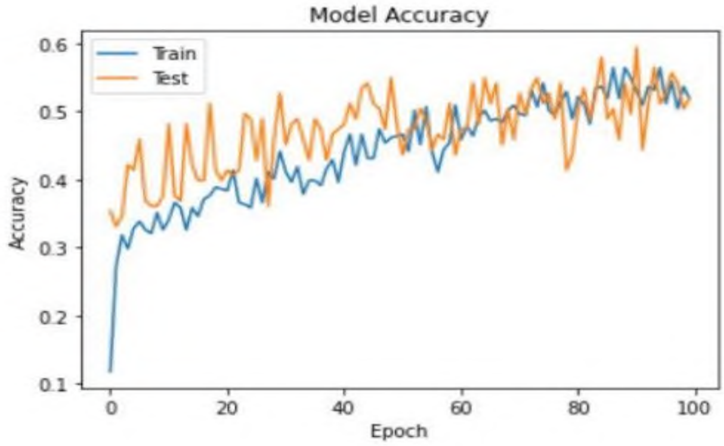


Figure 14. Model accuracy graph.

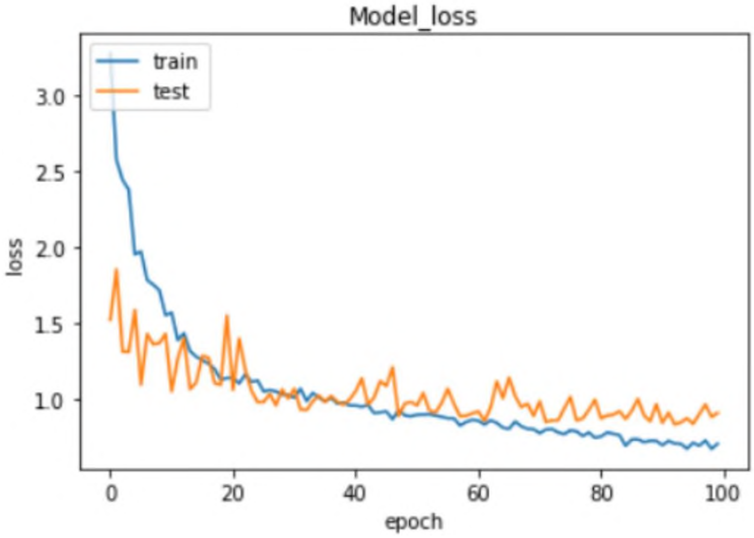


Figure 15. Model loss graph.

5. Conclusion

The main objective is to predict the emotional state of a customer from a short voice recording. The proposed model achieved 75.9% accuracy on testing data. The proposed approach is context independent; all audio segments were categorized separately. This work shows how deep learning can influence the fundamental feeling from audio speech data and a few insights on a person’s expression of feeling through voice. This scheme can be engaged in different environments like voice based virtual chatbots, Call Centre for marketing or complaints and linguistic research, etc.

Conflict of interest

The author declares no conflict of interest.

References

1. Morse NC. *Satisfactions in the White-Collar Job*. Survey Research Center; 1953.
2. Oliver RL. *Satisfaction: A Behavioral Perspective on the Customer*. McGrawHill; 1997.
3. Ali F, Ryu K, Hussain K. Influence of experiences on memories, satisfaction and behavioral intentions: A study of creative tourism. *Journal of Travel & Tourism Marketing* 2016; 33(1): 85–100. doi: 10.1080/10548408.2015.1038418
4. Chen CF, Chen FS. Experience quality, perceived value, satisfaction and behavioral intentions for heritage tourists. *Tourism Management* 2010; 31(1): 29–35. doi: 10.1016/j.tourman.2009.02.008
5. Khare SK, Bajaj V. An evolutionary optimized variational mode decomposition for emotion recognition. *IEEE Sensors Journal* 2021; 21(2): 2035–2042. doi: 10.1109/JSEN.2020.3020915
6. Wani TM, Gunawan TS, Qadri SAA, et al. A comprehensive review of speech emotion recognition systems. *IEEE Access* 2021; 9: 47795–47814. doi: 10.1109/ACCESS.2021.3068045
7. Mustaqeem, Sajjad M, Kwon S. Clustering-based speech emotion recognition by incorporating learned features and deep BiLSTM. *IEEE Access* 2020; 8: 79861–79875. doi: 10.1109/ACCESS.2020.2990405
8. Jiang P, Fu H, Tao H, et al. Parallelized convolutional recurrent neural network with spectral features for speech emotion recognition. *IEEE Access* 2019; 7: 90368–90377. doi: 10.1109/ACCESS.2019.2927384
9. Nwe TL, Foo SW, Silva LCD. Speech emotion recognition using hidden Markov models. *Speech Communication* 2003; 41(4): 603–623. doi: 10.1016/S0167-6393(03)00099-2
10. Wang J, Xue M, Culhane R, et al. Speech emotion recognition with dual-sequence LSTM architecture. *arXiv* 2019; arXiv:1910.08874. doi: 10.48550/arXiv.1910.08874
11. Akçay MB, Oğuz K. Speech emotion recognition: Emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers. *Speech Communication* 2020; 116: 56–76. doi: 10.1016/j.specom.2019.12.001
12. Chattopadhyay S, Dey A, Singh PK, et al. A feature selection model for speech emotion recognition using clustering-based population generation with hybrid of equilibrium optimizer and atom search optimization algorithm. *Multimedia Tools and Applications* 2023; 82(7): 9693–9726. doi: 10.1007/s11042-021-11839-3
13. Pandey SK, Shekhawat HS, Prasanna SRM. Multi-cultural speech emotion recognition using language and speaker cues. *Biomedical Signal Processing and Control* 2023; 83: 104679. doi: 10.1016/j.bspc.2023.104679
14. Singh P, Sahidullah M, Saha G. Modulation spectral features for speech emotion recognition using deep neural networks. *Speech Communication* 2023; 146: 53–69. doi: 10.1016/j.specom.2022.11.005
15. Pandey SK, Shekhawat HS, Prasanna SRM. Attention gated tensor neural network architectures for speech emotion recognition. *Biomedical Signal Processing and Control* 2022; 71, Part A: 103173. doi: 10.1016/j.bspc.2021.103173s