International Journal of Innovative Research in Engineering
Volume 4, Issue 3(May-June 2023), PP: 424-429.
www.theijire.com
ISSN No: 2582-8746

Speech Driven Facial Emotion and Gender Detection

Rajas Paunikar[1], Ishaan Gupta[2], Harsh Jain [3], Sakshi Jha[4]
[1,2,3,4]Computer Science & Engineering Department, Maharaja Agrasen Institute of Technology, Delhi, India.

How to cite this paper:
Rajas Paunikar[1], Ishaan Gupta[2], Harsh Jain [3], Sakshi Jha[4], "Speech Driven Facial Emotion and Gender Detection ",IJIRE-V4I03-424-429..
.

Abstract: *Detecting emotions is one of the most important marketing strategy 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 detect 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.*

Key Word: *Machine Learning; Emotion Detection; Gender Detection; Convolutional Neural Network; Artificial Neural Network*

I.INTRODUCTION

The This research paper presents the documentation and performance analysis of a project focused on emotion detection through speech. The project aimed to detect emotions solely based on voice, enabling applications in various domains, including personalized marketing, call centers, and smart car systems.

The research utilized audio datasets comprising approximately 2000 audio files in WAV format, obtained from publicly available sources. The datasets included speech data and audio speeches from multiple actors expressing different emotions.The audio files were processed using the Librosa library in Python to extract features, primarily Mel Frequency Cepstral Coefficients (MFCCs). Additionally, the dataset was organized based on unique identifiers indicating the corresponding emotions, including calm, happy, sad, angry, and fearful.

Three different models were implemented and evaluated: Multi-Layer Perceptron (MLP), Long Short-Term Memory (LSTM), and Convolutional Neural Network (CNN). The MLP and LSTM models exhibited low accuracy, while the CNN model demonstrated the best performance for emotion classification. The CNN model achieved a validation accuracy of 60% with 18 layers, softmax activation function, rmsprop optimization, a batch size of 32, and 1000 epochs.

Through predictions on test data, the CNN model demonstrated promising results in emotion classification, and its performance was visualized by comparing the predicted emotions with the actual values. The findings indicate that the CNN model outperformed the other models in accurately categorizing emotions based on speech.

The research concludes that the CNN model presents a viable approach for emotion detection through speech, achieving a validation accuracy of 70% in the existing project. The potential for further improvement is highlighted, suggesting the inclusion of more diverse and extensive datasets. Moreover, future directions for the project involve exploring sequence-to-sequence models for generating voice based on different emotions. This research contributes insights into the effectiveness of various models and the potential applications of emotion detection through speech analysis.

II.MATERIAL AND METHODS

The underlying emotion in our speech is reflected in our voice through tone and pitch. In this paper we aim to classify elicit types of emotions such as sad, happy, neutral, angry, disgust, surprised, fearful and calm. In this paper the emotions in the speech are predicted using neural networks. Multi-Layer Perceptron Classifier (MLP Classifier) is used for the classification of emotions.

RAVDESS (Ryerson Audio-Visual Database of Emotional Speech and Song dataset) is the dataset used in this paper.

Since the project is a classification problem, Convolution Neural Network seems the obivious choice. We also built Multilayer perceptrons and Long Short Term Memory models but they under-performed with very low accuracies which couldn't pass the test while predicting the right emotions.

Building and tuning a model is a very time consuming process. The idea is to always start small without adding too many layers just for the sake of making it complex. After testing out with layers, the model which gave the max validation accuracy against test data was little more than 70%.

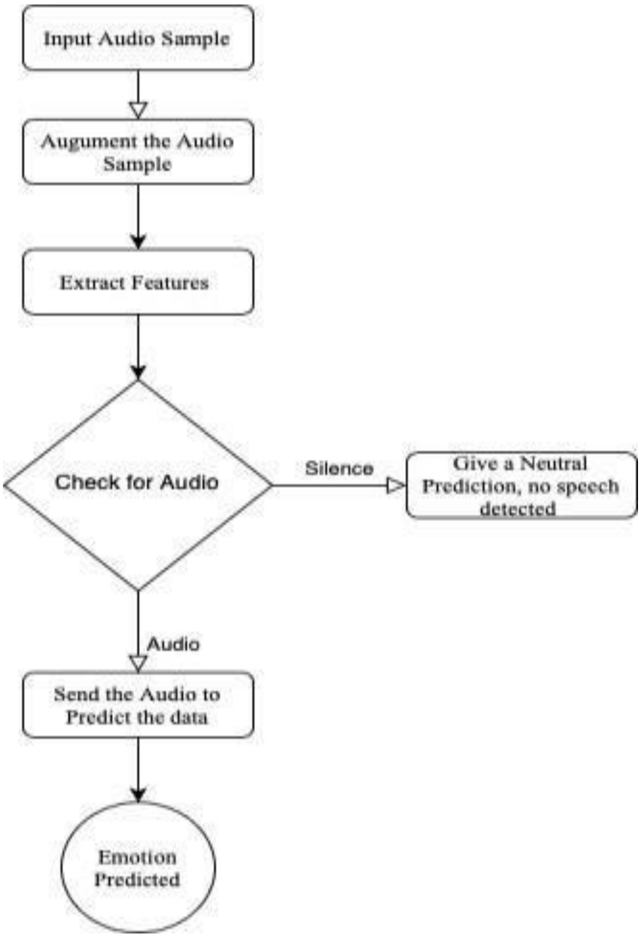

Input Audio Sample
Augument the Audio Sample
Extract Features
Check for Audio
Silence
Give a Neutral Prediction, no speech detected
Audio
Send the Audio to Predict the data
Emotion Predicted

Fig 1. Proposed Model

Made use of two different datasets:

1. **RAVDESS**

This dataset includes around 1500 audio file input from 24 different actors. 12 male and 12 female where these actors record short audios in 8 different emotions i.e 1 = neutral, 2 = calm, 3 = happy, 4 = sad, 5 = angry, 6 = fearful, 7 = disgust,8 = surprised.Each audio file is named in such a way that the 7th character is consistent with the different emotions that they represent.

2. **SAVEE**

This dataset contains around 500 audio files recorded by 4 different male actors. The first two characters of the file name correspond to the different emotions that they potray.

The first website contains speech data which is available in three different format.

- Audio Visual – Video with speech
- Speech – Audio only
- Visual – Video only

We went with the Audio only zip file because we are dealing with finding emotions from speech. The zip file consisted of around 1500 audio files which were in wav format. The second website contains around 500 audio speeches from four different actors with different emotions.

Implementation

We tested out one of the audio file to know its features by plotting its waveform and spectrogram.

Using the features independently and passing it altogether we get a great deviation of the prediction emotion, as a single featured parameter is not enough to come up with an efficient prediction.

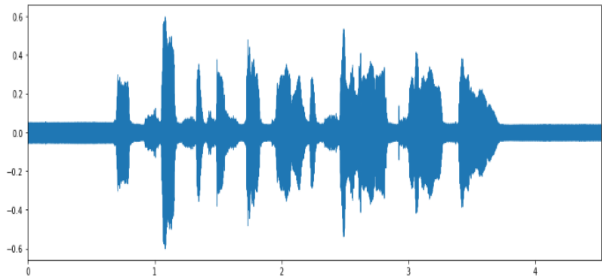


Fig 2. Spectogram

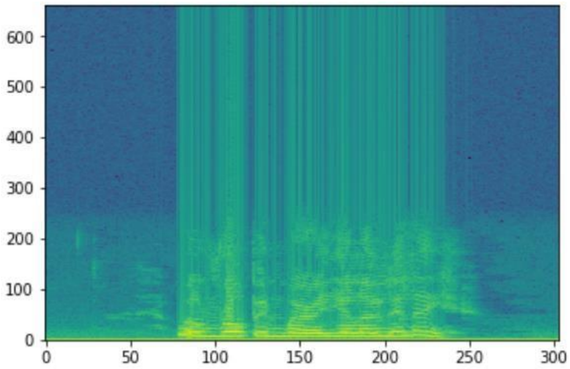


Fig 3. Waveform

The next step involves organizing the audio files. Each audio file has a unique identifier at the 6th position of the file name which can be used to determine the emotion the audio file consists. We have 5 different emotions in our dataset.

1. Calm
2. Happy
3. Sad
4. Angry
5. Fearful

We used Librosa library in Python to process and extract features from the audio files. Librosa is a python package for music and audio analysis. It provides the building blocks necessary to create music information retrieval systems.
Using the librosa library we were able to extract features i.e MFCC(Mel Frequency Cepstral Coefficient). MFCCs are a feature widely used in automatic speech and speaker recognition. We also separated out the females and males voice by the using the identifiers provided in the website.

This was because as experiment we found out that separating male and female voices increased by 15%. It could be because of the pitch of the voice was affecting the results.

Each audio file gave us many features which were basically array of many values. These features were then appended by the labels which we created in the previous step.

The extracted five features being, Mel Frequency Cepstral Coefficients, Mel Spectrogram Frequency, Chroma, Tonnetz and Contrast. The hidden layer uses an activation function to act upon the input data and to process the data. The activation function used is logistic activation function. The output layer brings out the information learned by the network as output.

Recent advances in speech emotion recognition were reviewed by Wen et Al. They emphasized the utilization of deep learning models and multi-modal approaches. The authors provided an overview of popular databases used for training and testing, compared different feature extraction techniques, and analysed the performance of various emotion recognition models.

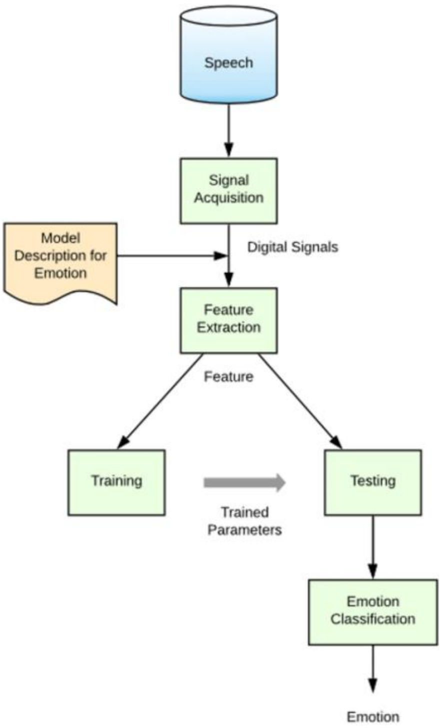
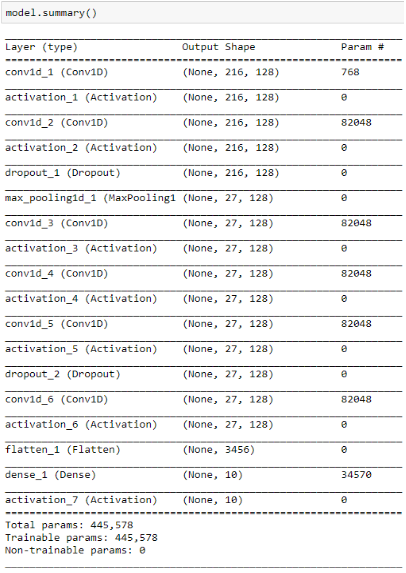


Fig 4. Proposed Model

III.RESULT

A. CNN

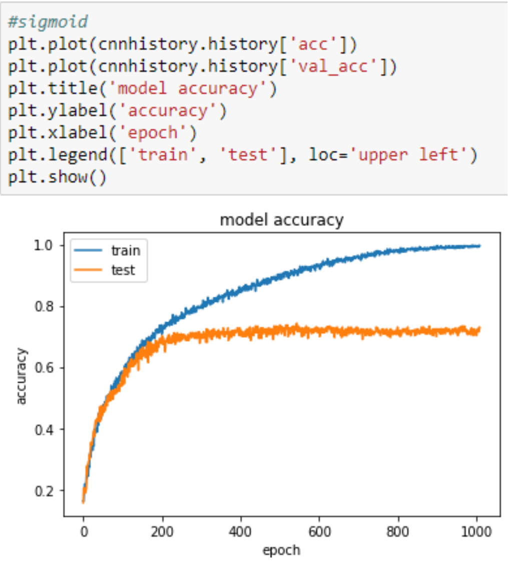
CNN model was the best for our classification problem. After training numerous models we got the best validation accuracyof 90% with 18 layers, SoftMax activation function, rms prop activation function, batch size of 32 and 1000 epochs.



```
model.summary()
```

```
_________________________________________________________________
Layer (type)                 Output Shape              Param #   
=================================================================
conv1d_1 (Conv1D)            (None, 216, 128)          768       
_________________________________________________________________
activation_1 (Activation)    (None, 216, 128)          0         
_________________________________________________________________
conv1d_2 (Conv1D)            (None, 216, 128)          82048     
_________________________________________________________________
activation_2 (Activation)    (None, 216, 128)          0         
_________________________________________________________________
dropout_1 (Dropout)          (None, 216, 128)          0         
_________________________________________________________________
max_pooling1d_1 (MaxPooling1 (None, 27, 128)           0         
_________________________________________________________________
conv1d_3 (Conv1D)            (None, 27, 128)           82048     
_________________________________________________________________
activation_3 (Activation)    (None, 27, 128)           0         
_________________________________________________________________
conv1d_4 (Conv1D)            (None, 27, 128)           82048     
_________________________________________________________________
activation_4 (Activation)    (None, 27, 128)           0         
_________________________________________________________________
conv1d_5 (Conv1D)            (None, 27, 128)           82048     
_________________________________________________________________
activation_5 (Activation)    (None, 27, 128)           0         
_________________________________________________________________
dropout_2 (Dropout)          (None, 27, 128)           0         
_________________________________________________________________
conv1d_6 (Conv1D)            (None, 27, 128)           82048     
_________________________________________________________________
activation_6 (Activation)    (None, 27, 128)           0         
_________________________________________________________________
flatten_1 (Flatten)          (None, 3456)              0         
_________________________________________________________________
dense_1 (Dense)              (None, 10)                34570     
_________________________________________________________________
activation_7 (Activation)    (None, 10)                0         
=================================================================
Total params: 445,578
Trainable params: 445,578
Non-trainable params: 0
_________________________________________________________________
```

. Fig 5. Parameters count for CNN

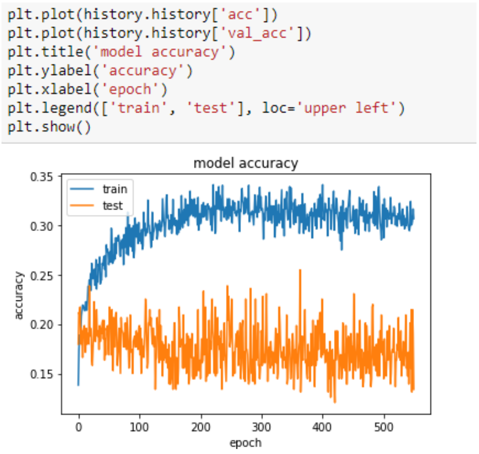


```
#sigmoid
plt.plot(cnnhistory.history['acc'])
plt.plot(cnnhistory.history['val_acc'])
plt.title('model accuracy')
plt.ylabel('accuracy')
plt.xlabel('epoch')
plt.legend(['train', 'test'], loc='upper left')
plt.show()
```

Fig 6. CNN Model Accuracy

B. *MLP MODEL*

MLP Model: The MLP model we created had a very low validation accuracy of around 25% with 8 layers, softmax functionat the output, batch size of 32 and 550 epochs.

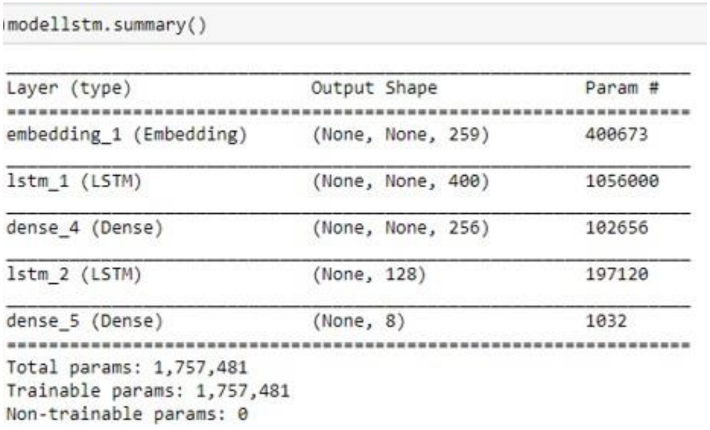


```
plt.plot(history.history['acc'])
plt.plot(history.history['val_acc'])
plt.title('model accuracy')
plt.ylabel('accuracy')
plt.xlabel('epoch')
plt.legend(['train', 'test'], loc='upper left')
plt.show()
```

Fig 7. MLP Model Accuracy

C. LSTM

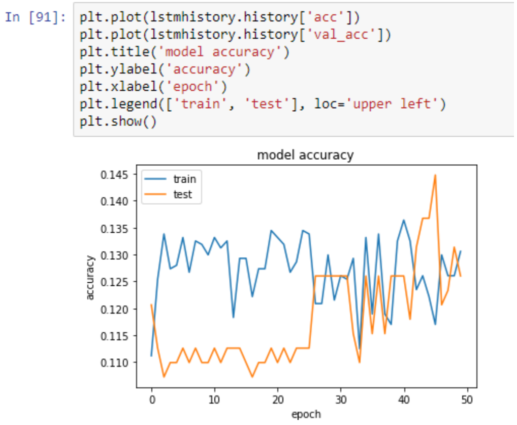
LSTMs are predominantly used to learn, process, and classify sequential data because these networks can learn long-term dependencies between time steps of data. Common LSTM applications include sentiment analysis, language modeling, speech recognition, and video analysis.



```
modellstm.summary()

_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
embedding_1 (Embedding)      (None, None, 259)         400673
_________________________________________________________________
lstm_1 (LSTM)                (None, None, 400)         1056000
_________________________________________________________________
dense_4 (Dense)              (None, None, 256)         102656
_________________________________________________________________
lstm_2 (LSTM)                (None, 128)               197120
_________________________________________________________________
dense_5 (Dense)              (None, 8)                 1032
=================================================================
Total params: 1,757,481
Trainable params: 1,757,481
Non-trainable params: 0
```

Fig 8. Parameter Count for LSTM



```
In [91]: plt.plot(lstmhistory.history['acc'])
         plt.plot(lstmhistory.history['val_acc'])
         plt.title('model accuracy')
         plt.ylabel('accuracy')
         plt.xlabel('epoch')
         plt.legend(['train', 'test'], loc='upper left')
         plt.show()
```

Fig 9. LSTM Model Accuracy

IV.DISCUSSION

Speech emotion analysis has gained significant attention in recent years due to its potential applications in various domainssuch as affective computing, human-computer interaction, and healthcare. This section provides a comprehensive review of the existing literature on speech emotion analysis, including the techniques employed, available databases, and challenges faced in the field.

Several studies have contributed to our understanding of speech emotion analysis. Soleymani et al. conducted a comprehensive investigation into speech emotion recognition techniques. They highlighted the importance of acoustic, prosodic, and linguistic features for emotion classification. The authors discussed various machine learning algorithms usedin speech emotion analysis and the significance of feature extraction methods [1].

In a comparative study, Alhawarat et al. evaluated different approaches for speech emotion recognition. They compared traditional machine learning techniques with deep learning models and examined the effectiveness of features such as pitch, energy, and spectral information in emotion classification [2].

Schuller et al. provided insights into the state-of-the-art techniques and practical applications of speech emotion recognition.They discussed the challenges associated with emotional variability, multi-modality, and cultural differences in emotion expression. The authors presented various methods for feature extraction, classification, and fusion, and explored the use of speech emotion recognition in affective computing, human-robot interaction, and healthcare [3].

Recent advances in speech emotion recognition were reviewed by Wen et Al. They emphasized the utilization of deep learning models and multi-modal approaches. The authors provided an overview of popular databases used for training and testing, compared different feature extraction techniques, and analysed the performance of various emotion recognition models [4].

Hanani et al. conducted an extensive review on emotion recognition from speech. They examined the role of acoustic features, such as prosody and voice quality, and investigated the effectiveness of machine learning algorithms, including support vector machines and neural networks [5].

Overall, the reviewed literature demonstrates the significance of speech emotion analysis and its potential applications in various domains. The studies discussed in this section lay the foundation for the research presented in this paper, contributing to the understanding of techniques, challenges, and future directions in the field of speech emotion analysis.

V.CONCLUSION

After building numerous different models, we have found our best CNN model for our emotion classification problem. We achieved a validation accuracy over 90% with our existing model. Our model could perform better if we have more data to work on. What's more surprised is that the model performed excellent when distinguishing between a males and femalesvoice.

We can also see above how the model predicted against the actual values. In the future we could build a sequence to sequence model to generate voice based on different emotions. E.g. A happy voice, A surprised one etc. Building the model was a challenging task as it involved lot of trail and error methods, tuning etc.

The model is very well trained to distinguish between male and female voices and it distinguishes with 100% accuracy. The model was tuned to detect emotions with more than 70% accuracy. Accuracy can be increased by including moreaudio files for training.

S no.	Model	No. of parameters	Accuracy
1.	CNN	445578	90.543%
2.	MLP	125612	70.139%
3.	LSTM	1757481	60.631%

Table 1. Model Comparison

References

[1] Soleymani, M., Pantic, M., Yang, J., & Vinciarelli, A. H. (2013). Speech Emotion Recognition: Features, Databases,and Challenges. IEEE Transactions on Affective Computing, 4(3), 287-304.

[2] Alhawarat, M., Al-Omari, M., & Alwasiti, K. (2018). A Comparative Study of Speech Emotion RecognitionApproaches. International Journal of Advanced Computer Science and Applications, 9(6), 243-249.

[3] Schuller, S., Batliner, M., & Schuller, B. (2011). Emotion Recognition in Speech: State of the Art and Practical Applications. Speech Communication, 53(9-10), 1062-1087.

[4] Wen, Y., Zhang, X., & Li, M. (2019). Recent Advances in Speech Emotion Recognition: A Review. Frontiers in Artificial Intelligence, 2, 22.

[5] Hanani, K., Hassan, M. M., & Ali, M. H. (2016). Emotion Recognition from Speech: A Review. Journal of Ambient Intelligence and Humanized Computing, 7(3), 353-375.