SPEECH BASED EMOTION RECOGNITION USING MACHINE LEARNING

[1.] G. Hrithik, [2.] K. Devi Kiran, [3.] K. N. V. Naresh,
[4.]G. Bharathi

Student, Department of CSE, SeshadriRaoGudlavalleruEngineeringCollege, Seshadri Rao Knowledge Village, Gudlavalleru, Andhra Pradesh, India
Sr. Gr. AssistantProfessor, Department of CSE, SeshadriRaoGudlavalleruEngineeringCollege, Seshadri Rao Knowledge Village, Gudlavalleru, Andhra Pradesh, India

hrithikgunduboyina@gmail.com
kiranreddy1738@gmail.com
mahinaresh412@gmail.com
gbharthi@gmail.com

.

ABSTRACT-**In intelligent speech applications, Speech Emotion Recognition (SER) is critical. The act of understanding human emotions and emotional states from speech is known as speech recognition. Human emotions can be expressed in a variety of ways, including bodily posture, face expression, and speech. Pitch, timbre, loudness, and vocal tone are some of the characteristics of the human voice. Humans have been found to express their emotions by changing voice characteristics during speech production. The suggested method uses Python modules such as PyAudio and Librosa for audio input and analysis of the audio, as well as the MLP Classifier for feature extraction and classification. The suggested method reports accuracy, f-score, precision, and recall for tested models in various experiment scenarios. However, earlier efforts have been hampered by a lack of attention to detail in the emotion prediction.**

1. INTRODUCTION

Speech emotion recognition (SER) is becoming more significant in a variety of applications like naturalistic human-computer interaction (HCI). Speech emotion recognition is now a popular study in the fields of signal processing and pattern recognition. The first part of a voice emotion processing and recognition system is signal collection, followed by feature extraction, and finally emotion recognition. The neural network-based approach is the most advantageous technology for speech recognition. Artificial Neural Networks (ANN) are information-processing technologies inspired by biological neural networks. Artificial neural networks (ANN) require no prior knowledge to simulate speech recognition.

2. LITERATURE SURVEY

Jianfeng Zhao et al.,proposedto understand speech emotion by analysing deep aspects from several data sources. To investigate the high-level points from unprocessed audio recordings and log-mel spectrograms, the

authors created a merged convolutional neural community (CNN) with two branches, one of which was a one-dimensional (1D) CNN branch and the other was a 2D CNN branch. There are two phases involved in creating the merged deep CNN. After designing and evaluating one 1D CNN and one 2D CNN architecture, the two CNN designs were fused together after removing the second dense layer. Switch learning was included in the training to help speed up the instruction of the combined CNN. The first to be trained were the 1D and 2D CNNs.Finally, the fine-tuning of the combined deep CNN initialised with transferred facets was performed.

S. Koelstra et al ,proposed a speech emotion detection model i.e a multimodal data set for the investigation of human affective states, was provided. While watching 40 one-minute long samples of music videos, the electroencephalogram (EEG) and peripheral physiological data of 32 subjects were monitored. Each film was scored on arousal, valence, like/dislike, dominance, and familiarity by the participants. The use of emotive tags offered as a unique way for stimuli selection. The results of a thorough examination of the participants' ratings throughout the experiment are presented. The researchers look into the correlations between EEG signal frequencies and participant ratings.

Babak Basharirad et al.,proposed a model due to the availability of high computation capability, the attention of the emotional speech signals research has been enhanced in human machine interfaces. Many approaches for detecting emotional states through speech have been proposed in the literature. The fundamental problems in speech emotion detection systems are the selection of adequate feature sets, the creation of proper classification methods, and the preparation of an appropriate dataset. Based on the three evaluation parameters, this research is critically examined the existing accessible techniques to speech emotion recognition algorithms (feature set, classification of features, accurate prediction). In addition, the performance and limits of available approaches are evaluated in this research. It also emphasises the present promising approach for voice emotion recognition system advancement.

3.PROPOSED WORK

Using the librosa, PyAudio modules and the Multi Layer Perception Classifier (MLP Classifier) and the RAVDESS Dataset, to create a model to discern emotion from speech. From sound files, the suggested system will be able to discern emotion. The data will be loaded, features extracted, and the dataset divided into training and testing sets. The model will then be trained using an MLP Classifier that has been initialised. Finally, we'll display the emotion detected from the user input audio. The suggested work achieves more accuracy and precision than existing systems, as well as the detection of more emotions than the prior system.

3.1 IMPLEMENTATION

3.1.1 MLP CLASSIFIER

A feedforward artificial neural network model, the Multi Layer Perceptron (MLP), maps input data sets to a collection of relevant outputs. An MLP is made up of numerous layers, out of which each is fully connected to the one before it.Multilayer Perceptrons are frequently trained on a set of input-output pairs and they are learned to model the correlation (or dependencies) between the input-output pairs.

.

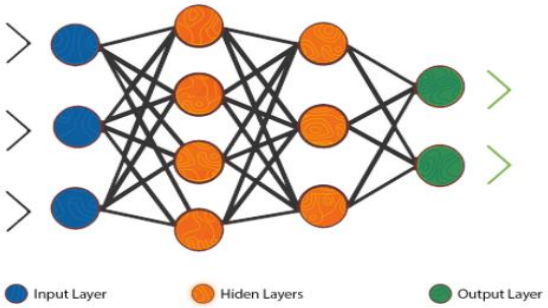

Fig 1:MLP Structure

3.1.2 ARCHITECTURE

The following diagram depicts the workflow of the proposed speech based emotion recognition system. At first, the audio is collected and pre-processed using the python library Librosa for removing the silence and noise reduction from the audio files. The extracted features from the RAVDESS dataset that undergoes training and testing process are stored based on their classification and a machine learning model is created. Now the input audio clip gets the feature extraction done and the features extracted are compared with the model that had undergo testing and training from the RAVDESS dataset. Finally, the output emotion is predicted and displayed on the screen.

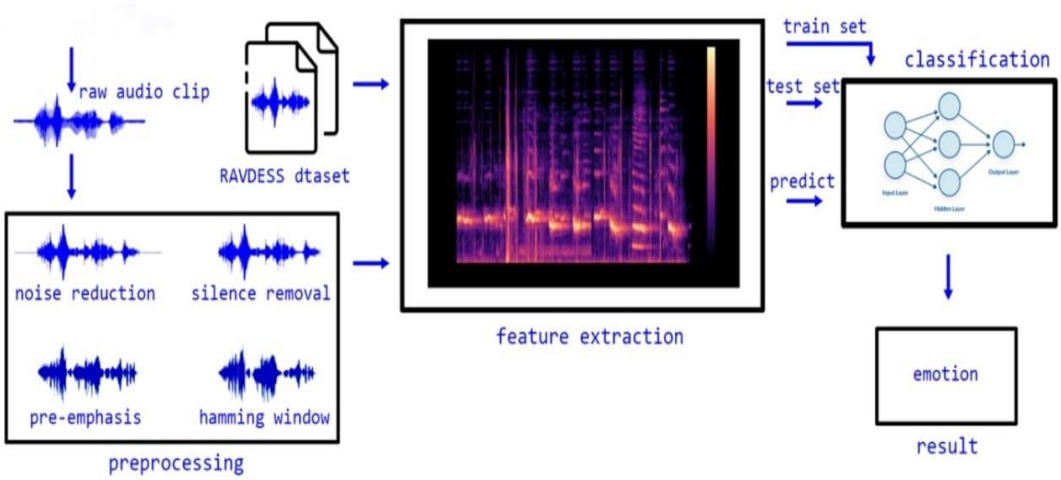

Fig 2:Work Flow of the proposed system

3.1.4. RAVDESS Dataset

The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) includes 7356 files, which is nearly 25GB of size. The database is said to contain 24 professional actors out of which 12 are female and 12 are male, vocalizing two lexically-matched statements.

The speech in the datasetincludes 8 emotions such as angry, fearful, sad, calm, neutral, disgust, happy and surpriseexpressions. Even though the database contains the songs and vidoes, since the proposed system is speech emotion detection we are taking only speech files as the dataset for testing and training of the model.

Speech folder contains 1440 files: 60 trials per actor x 24 actors = 1440. Each of the 7356 RAVDESS files has filename that is unique.(ex : 02-01-06-01-02-01-12.mp4).

Identifier	Coding description of factor levels
Modality	01 = Audio-video, 02 = Video-only, 03 = Audio-only
Channel	01 = Speech, 02 = Song
Emotion	01 = Neutral, 02 = Calm, 03 = Happy, 04 = Sad, 05 = Angry, 06 = Fearful, 07 = Disgust, 08 = Surprised
Intensity	01 = Normal, 02 = Strong
Statement	01 = "Kids are talking by the door", 02 = "Dogs are sitting by the door"
Repetition	01 = First repetition, 02 = Second repetition
Actor	01 = First actor, . . ., 24 = Twenty-fourth actor

We have get this dataset from thekaggle

https://www.kaggle.com/uwrfkaggler/ravdess-emotional-speech-audio#:~:text=Ryerson%20Audio%2DVisual,number%20is%20even

3.1.5 EMOTIONS DETECTING :

The following emotions will be predicted in the proposed speech based emotion recognition system. The proposed system also returns the gender of the user input audio file that is sent for detecting the emotion.

The following figure indicates the name of the file for predicting the audio for classification in the model building(i.e if the file name contains '01' indicates the audio file is neutral emotion, which are required during the testing and training process of model building).

```
emotions={
  '01':'neutral',
  '02':'calm',
  '03':'happy',
  '04':'sad',
  '05':'angry',
  '06':'fearful',
  '07':'disgust',
  '08':'surprised'
}

def gender(g):
    if int(g[0:2]) % 2 == 0:
        return 'female'
    else:
        return 'male'
```

4 FEATURE EXTRACTION

The feature extraction process extracts the features from the given dataset. However, the features extracted will be different from each other based on the file type(i.e the feature extraction for an image file will be different from the feature extraction for the audio file).

The proposed system is based on audio and there are many features to determine the audio classification. But for the proposed system these are the features used to classify the audio.

Audio Features :

Mfcc : Mel Frequency Cepstral Coefficient, defines the sound's short-term power spectrum. It entails windowing the signal, calculating the log of the magnitude, wrapping the features on a mel scale, and also performing the inverse DCT.

Chroma : Pertains to the 12 different pitch classes. High-level features intend to build user related concepts out of low and mid level features.

Mel : Mel Spectrogram Frequency. It defines the pitches judged by listeners to check the distance to be equal from each another.

5.RESULTS AND DISCUSSION

5.1 Features Extracted :

The following depicts the total number of features extracted from the dataset used for training the model. The feature extraction resulted by extracting 180 features from the dataset given for training the model.

```
print((X_train.shape[0], X_test.shape[0]))
print(f'Features extracted: {X_train.shape[1]}')
```

```
(1152, 288)
Features extracted: 180
```

5.2 Classification Report & Confusion Matrix :

Classification report is often defined as the performance metric for the machine learning model. It is used to determine the precision, support, f1-score and recall of the machine learning model. The following depicts the support, precision, f1-score and recall of the emotions (angry, fearful, sad, calm, neutral, disgust, happy and surprise expressions) of the built machine learning model.

	precision	recall	f1-score	support
angry	0.71	0.69	0.70	52
calm	0.67	0.15	0.24	40
disgust	0.25	0.71	0.37	48
fearful	0.49	0.69	0.57	36
happy	0.55	0.36	0.44	47
neutral	0.12	0.03	0.05	29
sad	0.36	0.33	0.34	52
surprised	0.75	0.38	0.50	56
accuracy			0.44	360
macro avg	0.49	0.42	0.40	360
weighted avg	0.51	0.44	0.42	360

The confusion matrix is also one of the performance metric for evaluating the efficiency of the model. The following is the confusion matrix of the model of the given 8 (i.e 8*8 matrix) emotions.

```
[[36  0 10  1  2  0  2  1]
 [ 0  6 19  5  0  3  7  0]
 [ 3  0 34  3  4  1  2  1]
 [ 4  0  3 25  0  1  3  0]
 [ 6  0 13  2 17  2  4  3]
 [ 0  0 13  4  0  1 10  1]
 [ 1  3 19 10  1  0 17  1]
 [ 1  0 24  1  7  0  2 21]]
```

5.3 Accuracy:

The accuracy of the model is the ultimate metric in evaluating the performance of the machine learning model. The accuracy of the proposedmodel using MLP(Multi Layer Perceptron) Classifier along with the LGB(Light Gradient Boost) combined by using the voting classifier resulted in higher accuracy compared to the previous systems.

```
for key in models.keys():
    fmodel=model(key,save=True,print_stat=False,cv=True)
```

```
Voting Classifier: MLP2, LGB CV Accuracy: 0.7013739883305101
```

5.4 Output:

The user input audio file should be specified in the predict function cell to predict the emotion of the speech and gender of the audio, along with the text spoken in the audio and they are displayed on the screen.

```
predict("/content/drive/MyDrive/Main Project/speech/Actor_14/03-01-01-01-01-01-14.wav")
```

```
['neutral_female', 'kids are talking by the door']
```

```
predict("/content/drive/MyDrive/Main Project/test.wav")
```

```
['surprised_female', 'dogs are sitting by the door']
```

```
predict("/content/drive/MyDrive/Main Project/Monday at 3-07 pm.wav")
```

```
['fearful_male', 'I am very afraid']
```

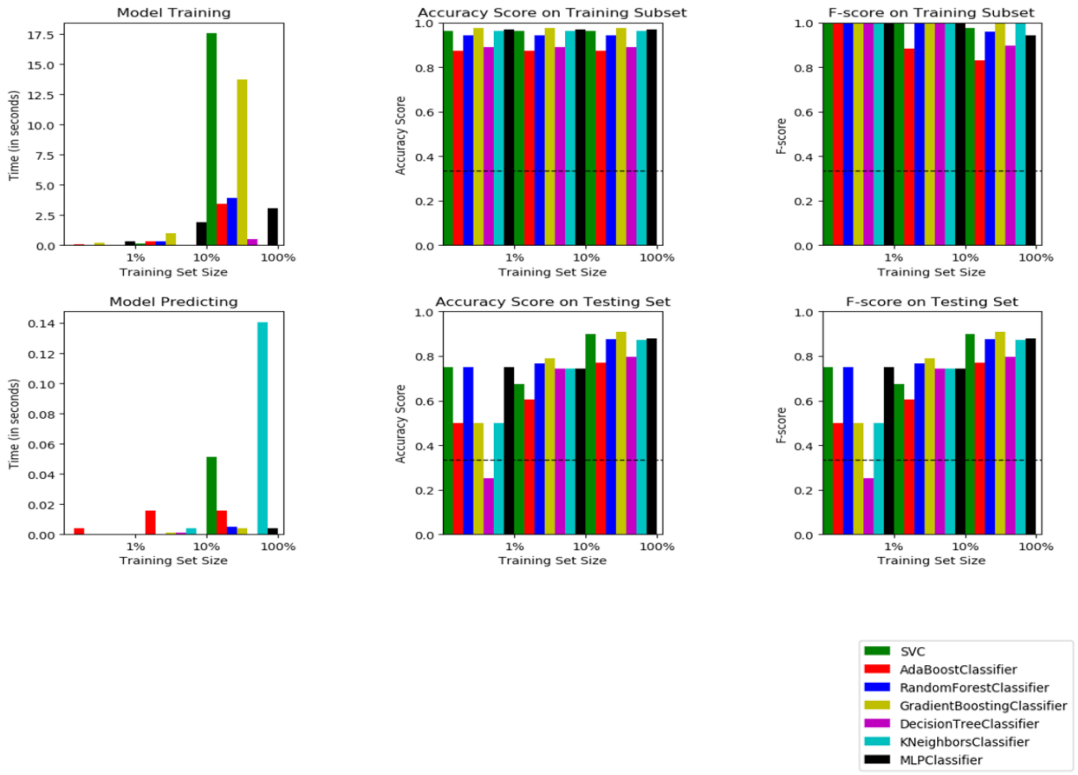

Fig 3 : Study on different Classifiers and their accuracies in Detecting emotion from speech.

The above graph depicts the time taken for predicting the emotion, accuracy scores and f-scores of both the testing set and training set of different classifiers in predicting the emotion from speech. The lesser the time taken for prediction indicates the better efficiency of the model. Some classifiers produced better results of accuracy scores and some classifiers produced better results of f-scores. However, the classifiers that produced greater results in accuracy and f-scores are taking more time in the model prediction. We observed that the MLP Classifier performs consistently better in both the accuracy and f-scores and also predicts the model in less time compared to other classifiers.

6.CONCLUSION

The paper indicates based on the proposed work that the MLPs are very accurate and effective in classifying and predicting emotion from the speech signals. Even with limited models, a restricted set of features can be extracted without problems identified. We have acquired greater accuracies as in contrast to different techniques for character emotions. The accuracy of the proposed model is 70%. The overall performance of the model is pretty structured on the great of pre-processing. Almost, every human emotion required for the proposed system, has been completely studied, analyzed and the accuracy has been checked.The consequences received in this learn about display that speech focus is feasible, and that MLPs can be used in recognizing the speech and depicting the accuracy of every emotion existing in the speech.

REFERENCES

[1] TuomasEerola and Jonna K. Vuoskoski, “A comparison of the discrete and dimensional models of emotion in music”, Psychology of Music, pp. 1–32, The Author(s) 2010.

[2] Eddie Harmon-Jones, Cindy Harmon-Jones, and Elizabeth Summerell, “On the Importance of Both Dimensional and Discrete Models of Emotion” School of Psychology, The University of New South Wales, Australia.

[3] Roddy Cowie, Ellen Douglas-Cowie, Susie Savvidou, Edelle McMahon, Martin Sawey& Marc Schröder, ‘FEELTRACE’ Schools of Psychology and English, Queen's University Belfast.

[4] S. Koelstra, C. Muhl, M. Soleymani, J.-S. Lee, A. Yazdani, T. Ebrahimi, T. Pun, A. Nijholt, and I. Patras, “Deap: A database for emotion analysis using physiological signals,” IEEE Transactions on Affective Computing, vol. 3, no. 1, pp. 18–30, 2012.

[5] “Using Deep and Convolutional Neural Networks for Accurate Emotion Classification on DEAP Dataset,” Proceedings of the Twenty-Ninth AAAI Conference on Innovative Applications(IAAI-17)).

[6] Chung, SeongYoub, and Hyun Joong Yoon. ”Affective classification using Bayesian classifier and supervised learning.”Control,Automation and Systems (ICCAS), 2012 12th International Conference on. IEEE, (2012).

[7] Jaebok Kim, Khiet P. Truong, Gwenn Englebienne, and Vanessa Evers, “Learning spectro-temporal features with 3D CNNs for speech emotion recognition“,Human Media Interaction, University of Twente, Enschede, The Netherlands, 2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII).

[8] Babak Basharirad, and Mohammadreza Moradhaseli,“Speech emotion recognition methods: Literature review”,A IP Conference Proceedings 1891, 020105 (2017).

[9] Kim, Jaebok and Englebienne, Gwen and Truong, Khiet P and Evers, Vanessa, ”Towards Speech Emotion Recognition ``in the wild" using Aggregated Corpora and Deep Multi-Task Learning”, ’Proceedings of the INTERSPEECH’, year-2017.

[10] B. Yang, M. Lugger, “Psychological motivated multistage emotion classification exploiting voice quality feature.” F. Mihelic, J. Zibert, Speech Recognition, InTech, 2008, chapter 22.

[11] C.-H. Wu and W.-B. Liang, “Emotion Recognition of Affective Speech Based on Multiple Classifiers Using Acoustic-Prosodic Information and Semantic Labels,” IEEE Trans. Affect. Comput., vol. 2, no. 1, pp. 5–20, Jan. 2011.

[12] Steven R. Livingstone, 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”,journal.pone.0196391,May 16, 2018.

[13] Monorama Swain, AurobindaRoutray, PrithvirajKabisatpathy,"Databases, features and classifiers for speech emotion recognition: a review", I. J. Speech Technology 2018.

[14] S. G. Koolagudi and K. S. Rao, "Emotion recognition from speech: a review," International Journal of Speech Technology, vol. 15, no. 2, pp. 84–115, 2012.