iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 8 **Issue:** V **Month of publication:** May 2020

DOI: http://doi.org/10.22214/ijraset.2020.5354

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 8 Issue V May 2020- Available at www.ijraset.com

Speech Emotion Identification

Yash Desai[1], Yashowardhan Rungta[2], Ashwati Iyer[3], Sarthak Chandarana[4]
[1, 2, 3, 4]*Department of Electronics and Telecommunication, NMIMS-Mukesh Patel school of Technology Management and Engineering, Mumbai,India*

Abstract: This paper is an effort at developing a Speech Emotion Identifier model by implementing Librosa and sklearn libraries on a RAVDESS dataset. It gives the reader an insight on the way to detect a human's emotion based on the speech, taken as input audio file. A newly developed speech signal model is applied to provide the user with the likelihood that the given speech is a response to a given emotion.
This particular model is built using convolution neural networks (CNN) and classifiers namely Decission Tree, Random Forest and Multi-Layer Perceptron This model finds its applications in various real-world scenarios and therefore the most potent example for the identical would be in Customer care services where the staffs keep changing their way of pitching by recognizing Customers' emotion from their speech so as to improve their quality of services provided. This paper presents the feasibility of extraction of MFCC features within the model.
This model takes into consideration three different classifiers MLP, Random Forest and Decision Tree and by taking a combination of these three, we get the best possible accuracy as output.
Keywords: Librosa, sklearn, MLP Classifier, Random Forest Classifier, Decision Tree Classifier, MFCC, CNN

I. INTORDUCTION

First, Communication is a very important part of understanding various human beings. It is the basic mode of interaction between various individuals. At times, we hide our real emotions behind the veil of words that we speak. Understanding the emotion of the speaker from his / her speech is a complex task.
This often can lead to erroneous assumptions and conclusions. Mankind has always tried to evolve and invent various technologies challenging himself to the utmost level of his mind to innovate and present newer established technologies. One such on-going development model is the Speech Emotion Recognizer (SER).
Contributions from several programmers to this relatively new field of Research have been done and it is still a work in progress. In such a model that has wide applications in various fields, complexity of implementation does knock the door as, if imagine, humans themselves cannot completely understand the emotions behind the speech then how could one expect a virtual interface to do the same? Thus, this model takes up these challenges and delivers the best result possible.

II. LITERATURE SURVEY

Speech Emotion Recognition is one of the most challenging tasks in the speech analysis domain. This paper provides us with highest accuracy amongst many papers from the past as it is an integration of multiple classifiers such as MLP, Random Forest and Decision Tree Classifier.
The motivation for this paper has been taken from research papers published earlier. These papers provide us with an insight to how SER can be used in the fields of household, military as well as medical advancement.[2]
It got simpler to recognize the enthusiastic states in vocal articulations by removing speech highlights from speech datasets utilizing ML algorithms and it gave a concise outline of the present situation. [4].
The use of the Inception Net model along with deep learning techniques to recognize and classify the emotion into a list [2] helped in classifying the emotions into a list from 0-7, making it easier to observe the test data set.
This paper is an integration of multiple classifiers powered by an extensive dataset which makes it more realistic and technologically diverse. This technological diversity provides us with increased accuracy and the different sources of audio samples prevent this model from becoming monotonous. Models in the past have been accurate to around 35%. However, this model that we have used is accurate up to 60%.

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 8 Issue V May 2020- Available at www.ijraset.com

III. DATASET AND ALGORITHM

This section gives the reader an in-depth detail of the dataset used, the classifiers implemented and the features extracted for creation of the proposed model. It also explains about Convolution Neural Network (CNN) and how the design of this model was plotted.

A. Dataset Extraction and Sample Naming

In this model, we have used a Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) Dataset which is useful for implementation in Machine Learning Projects. The original dataset consisted of various types of files like audio, video, speech audio-only files corresponding to the Eight emotions that we have considered. Each of the eight emotions is produced at 2 different intensity levels (normal, strong). We segregated the speech audio-only files for various emotions and extracted it from the original 24.8 GHz dataset to form the model's dataset consisting of 1479 speech audio-only samples. Each of the sample filename is unique and consists of a Seven-part numerical identifier. The numbering of each filename identifier is done as follows:

1) *Modality:* The first prefix of the file name depicts modality of the file and predicts whether the file is an audio only, video only or audio-video file. Only audio files have been extracted and used for this model
2) *Emotion:* The third prefix of the file name depicts the emotion of the speech and predicts various emotions like neutral, calm, happy, sad etc. in a numerical order 01-08 respectively.
3) *Actor:* The last prefix of the file name depicts the gender of person speaking where even numbers represent female and odd numbers represent male.

We have considered audio-only samples and thus all the sample filenames have the start prefix: 03-01.
For example: "03-01-05-01-01-02-02.wav" is an audio-only speech sample portraying angry emotion in normal intensity repeating "Kids are talking by the door" statement twice. The actor speaking is a female [1]. The classified emotions according to the dataset are 1-8 but in our extracted dataset, for ease of computation, we have changed the same from 0-7 using simple for loop.

B. Classification and Algorithm

Classification is approximating a mapping function (f) from input variables (x) to predicted output variables(y).
There are various classification algorithms available and in a generalized scenario, it is difficult to quote which one is the best/is better than the other.
Based on the application of these for specific dataset used, it can be determined which one is superior than the other algorithm used for that particular dataset.
For this particular model, we took into consideration three different classifiers:

1) *MLP Classifier:* Multi-layer Perceptron classifier connects itself to a Neural Network and relies on it for classification of dataset. Its implementation from sklearn can be done effortlessly.
2) *Random Forest:* A gathering learning technique for grouping, relapse and different errands that works by developing various choice trees at preparing time and yielding the class that is the method of classes (arrangement) or mean forecast (relapse) of the individual trees. It revises for choice trees' propensity for over-fitting to their training set.
3) *Decision Tree:* It is a notable grouping strategy in various example acknowledgment issues like, picture order and character acknowledgment. They perform all the more effectively, explicitly for complex grouping issues, because of their high flexibility and computationally successful highlights. Furthermore, it likewise surpasses desires over various ordinary regulated characterization techniques.

This model is a combination of all these three classifiers which provide 60% testing accuracy

C. MFCC feature

Mel-frequency cepstral coefficients will be coefficients that by and large make up a MFC which is the Mel-frequency cepstrum that speaks to a transient power spectrum of a sample, in light of a direct cosine change of a log power range on a nonlinear mel scale of frequency. They are gotten from a kind of cepstral portrayal of the audio suppression.
In Mel-frequency cepstrum, the frequency groups are similarly divided on the mel scale and it approximates the human sound-related framework's reaction more intently than the typical cepstrum frequency groups. This frequency distortion can take into consideration better portrayal of sound, for instance, in sound compression.
In this model, for each of the sample audio file, 40 MFCCs have been extracted that is 1479x40=59160 MFCCs have been extracted.

D. Convolution Neural Network

In deep learning, it is a class of acute neural systems, ordinarily utilized for breaking down visual symbolism.

Convolution is basically used to find key features from image using feature detector. A feature map is created that preserves the spatial relationship between pixels.

CNN is a numerical build that is regularly made out of three kinds of layers: convolution, pooling, and completely associated layers. The initial two, convolution and pooling layers, perform highlight extraction, though the third one is a completely associated layer, that maps the extricated highlights into conclusive yield like arrangement .The input layer of our model containsall the audio samples while the output layer has all the eight classified emotions ranging from 0-7 (as mentioned in 3.1).

Conv 1d is implemented as the dataset consists of 59160 MFCCs extracted and the features are located at random locations. Hence, to analyse a particular feature

whose location is not of concern, of any audio sample extracted for a fixed-length period, 1d CNN is used.

The prediction of each emotion is done at the final layer but selection of a particular emotion from various hidden layers is done at the Activation layer.

The type of activation function used in this model is ReLU which stands for Rectified linear unit and mathematically defined as y=max (0,x). It gives a linear output for all positive values and zero for negative values. It classifies the input audio sample based on its maximum weighted sum and determines the category of emotion as listed in the 3.1

The training data set consists of 1109 samples and when it is tested, it gives us a very high accuracy i.e. 93%. However when the model is applied to a new unseen dataset i.e. testing dataset, the accuracy drops. This is called over-fitting. To tackle this predicament, we are using the Drop-Out function. Drop-Out is a regularization technique which has been patented by Google Inc. This technique drops out unnecessary random neurons during activation to make the model more compact and hence providing a higher testing accuracy

IV. EXPERIMENTAL RESULTS

The input audio samples were loaded using Librosa library's load function. The sampling rate for these samples was 22 KHz i.e. the Nyquist rate, which is half the rate at which humans communicate i.e. 44 KHz

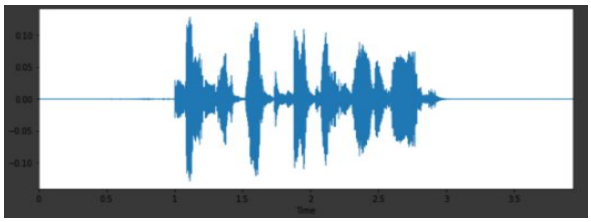


Fig. 1 Waveform of the input audio sample

The waveform describes a depiction of the pattern of sound pressure variation (or amplitude) in the time domain.

After loading the dataset, the audio sample's features(MFCC) of the entire dataset are extracted. A total of 40 MFCCs are extracted for each file on the dataset. The loaded dataset is divided into training(75%) and testing(25%) dataset These features are then loaded into different classifiers; namely Random Forest, MLP, Decision Tree. The classification reports of each of them are given below:

	precision	recall	f1-score	support
0	0.67	0.08	0.14	25
1	0.47	0.91	0.62	46
2	0.42	0.22	0.29	51
3	0.47	0.45	0.46	42
4	0.58	0.65	0.61	51
5	0.50	0.49	0.49	39
6	0.55	0.46	0.50	61
7	0.59	0.71	0.64	55
accuracy			0.52	370
macro avg	0.53	0.50	0.47	370
weighted avg	0.53	0.52	0.49	370

(a)

	precision	recall	f1-score	support
0	0.00	0.00	0.00	25
1	0.43	0.65	0.52	46
2	0.40	0.16	0.23	51
3	0.26	0.76	0.39	42
4	0.61	0.67	0.64	51
5	0.73	0.21	0.32	39
6	0.53	0.26	0.35	61
7	0.48	0.55	0.51	55
accuracy			0.43	370
macro avg	0.43	0.41	0.37	370
weighted avg	0.46	0.43	0.40	370

(b)

	precision	recall	f1-score	support
0	0.39	0.36	0.37	25
1	0.64	0.54	0.59	46
2	0.29	0.25	0.27	51
3	0.23	0.24	0.23	42
4	0.45	0.63	0.52	51
5	0.26	0.31	0.28	39
6	0.34	0.31	0.32	61
7	0.44	0.36	0.40	55
accuracy			0.38	370
macro avg	0.38	0.38	0.37	370
weighted avg	0.38	0.38	0.38	370

(c)

Fig. 2 (a) Random forest, (b) MLP, (c) Decision tree

Layer (type)	Output Shape	Param #
conv1d_1 (Conv1D)	(None, 40, 128)	768
activation_1 (Activation)	(None, 40, 128)	0
dropout_1 (Dropout)	(None, 40, 128)	0
max_pooling1d_1 (MaxPooling1	(None, 5, 128)	0
conv1d_2 (Conv1D)	(None, 5, 128)	82048
activation_2 (Activation)	(None, 5, 128)	0
dropout_2 (Dropout)	(None, 5, 128)	0
flatten_1 (Flatten)	(None, 640)	0
dense_1 (Dense)	(None, 8)	5128
activation_3 (Activation)	(None, 8)	0

Total params: 87,944
Trainable params: 87,944
Non-trainable params: 0

Fig. 3 Model summary of the Neural Network

After training this model, we obtain training accuracy of 92.34% and testing accuracy of 59.46%.

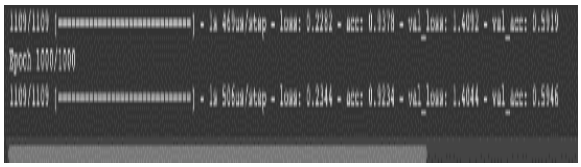


Fig. 4 Training and testing accuracy of the model

This difference in training and testing accuracy can be overcome by reducing epochs.

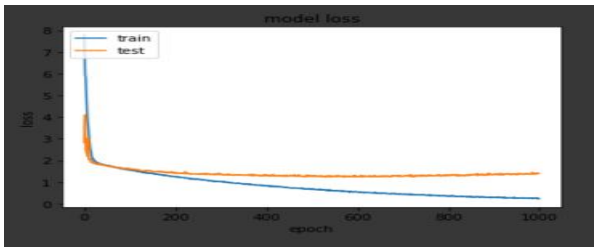


Fig 5. Model trained for 1000 epochs

It is visible from the plot that there is a small amount of over fitting present in the model. Hence, to obtain an efficient model, we train the model again for 250 epochs.

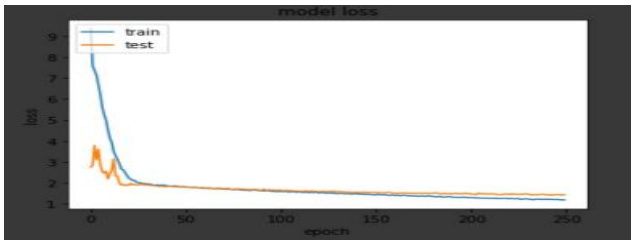


Fig. 6 Model trained for 250 epochs

We performed training of 1109 audio samples and tested the remaining 370 audio samples of the dataset. The Image shown below gives us the prediction of every single emotion of the testing dataset

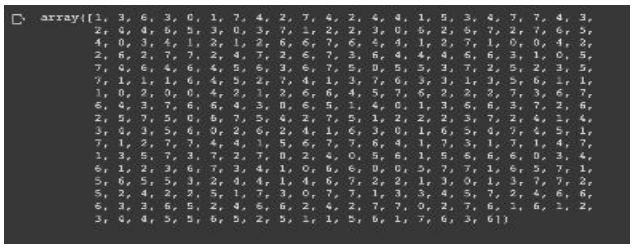


Fig. 7 testing the model

The accuracy obtained of the model is 59.46% which is higher compared to existing models.

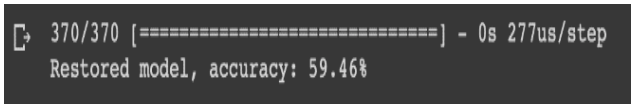


Fig. 8 Accuracy of the designed model

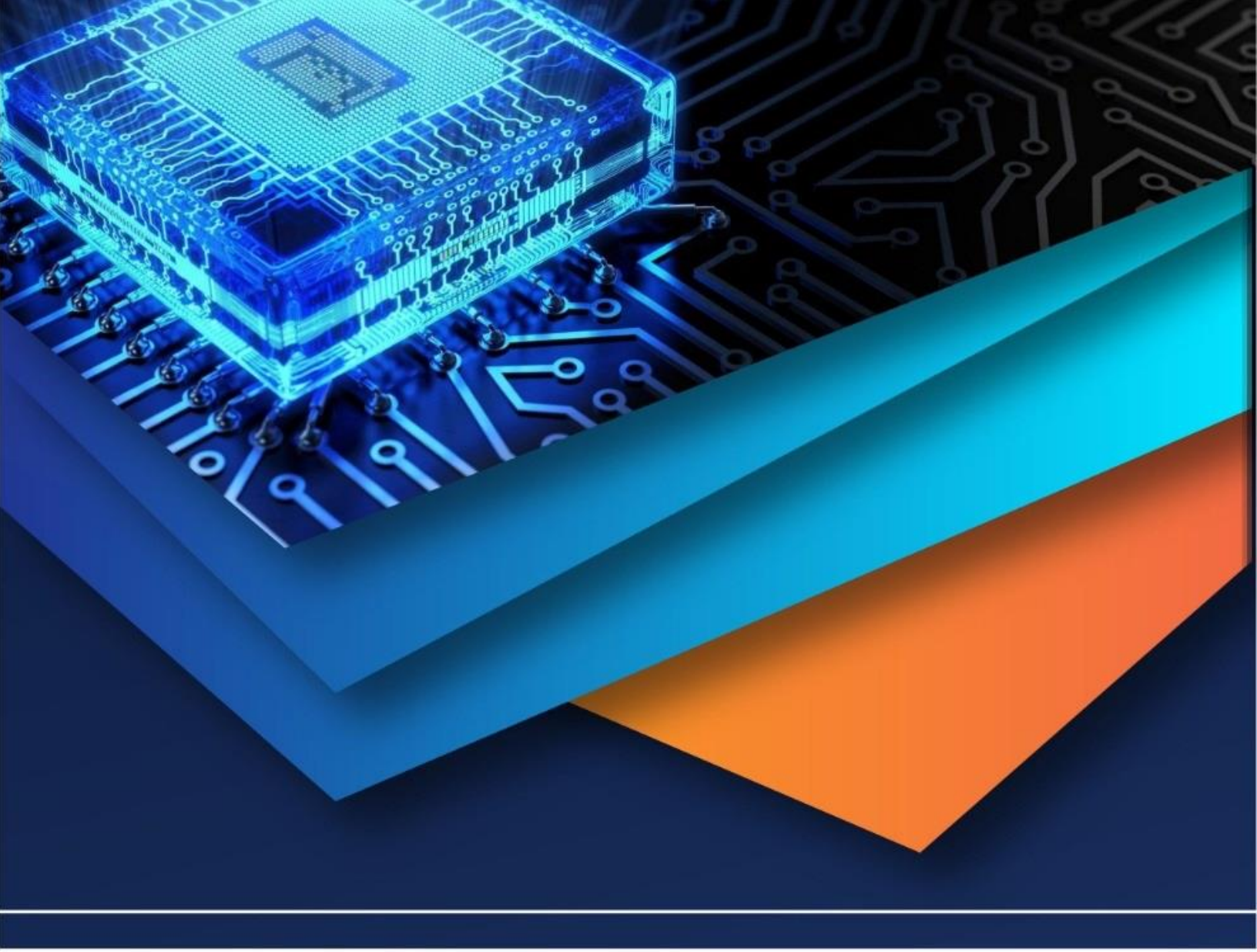
Once the trained model is obtained, we can extract the features of a particular input audio sample and pass them from the model to predict the emotion of that audio sample. This can be done simply, by using the trained model and the Predict function.

V. CONCLUSION

The accuracy of the model is approximately 60% and it can be used in medical, military as well as household applications to detect sentiment amongst the human race. The accuracy of the model can be improved by using a dataset with increased number of samples, which in turn, reduces the over-fitting problem too. A greater number of features can also be extracted in order to obtain comparatively precise classifier outputs. This model can also be implemented for various other languages other than English. This could improve its scope for various applications worldwide. It can also be used by psychiatrists for getting a better idea about their patients' feelings and emotions more effectively.

REFERENCES

[1] Livingstone SR, Russo FA (2018), "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(5): e0196391. https://doi.org/10.1371/journal.pone.0196391.
[2] S. Lugović, I. Dunđer and M. Horvat, "Techniques and Applications of Emotion Recognition in Speech (2016)": https://ieeexplore.ieee.org/document/7522336
[3] Nithya Roopa, S. Prabhakaran and M,Betty, "Speech Emotion Recognition using Deep Learning (2018)": https://www.ijrte.org/wpcontent/uploads/papers/v7i4s/E1917017519.pdf
[4] Xu Huahu, Gao Jue and Yuan Jian, "Application of Speech Emotion Recognition in Intelligent Household Robot (2010)": https://ieeexplore.ieee.org/document/5655398
[5] C. Busso, A. Metallinou, and S. S. Narayanan, "Iterative feature normalization for emotional speech detection," in Proceedings of IEEE ICASSP 2011. IEEE, 2011, pp. 5692–5695.
[6] A. Graves, A. Mohamed, and G. Hinton, "Speech Recognition with Deep Recurrent Neural Networks," in 2013 IEEE International Conference on Acoustics, Speech and Signal Processing, 2013, pp. 6645–6649.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY