International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 07 Issue: 11 | Nov 2020 www.irjet.net p-ISSN: 2395-0072

Speech Emotion Recognition using CNN

Sagar Borkar[1], Urvi Shukla[2], Naman Bangad[3], Aashutosh Awasthi[4], Pratik Dabre[5]

[1,4]Student, Computer Engineering Department, Thakur College of Engineering and Technology, Maharashtra, India.
[5]Student, Computer Engineering Department, Fr Conceicao Rodrigues College of Engineering, Maharashtra, India.

---***---

Abstract - *Given the Modern-day scenario, where Artificial intelligence equipped Voice Assistants play a major role in every household, understanding the speech sentiments of the person becomes utmost important. Speech Emotion Recognition also plays an important role while a customer is interacting with a customer care service irrespective of customer care being a person or an automated machine. In this Machine learning based project of ours, the data which we have used is from 4 sources – CREMA-D, RAVDESS, SAVEE, TESS coupled with Python based library - Librosa and keras for convolutional*

Neural networks. The data is splitted into 3:1 ratio as Train data and Test data. We have used Sequential modeling. We have used wave plot and spectrogram for audio data analysis and visualization along with subplot and confusion matrix.

Keywords: **Machine learning, CNN, Sentiment analysis, SER, HCI.**

1. INTRODUCTION

Humans communicate through many ways but verbal communication is the easier way of communication. Verbal communication takes place through speech and language. The way of delivering words says almost everything about the emotion of humans. SER (Speech Emotion Recognition) is the very powerful attempting act to understand the Emotion of humans and help us to recognize affective states from the speech. This is the true fact that voice reflects our emotion through tone and the pitch. That's the reason by which animals are easily able to understand the emotions of the Human. It is the very recent research topic in the Human computer interaction (HCI). As the computer became a very integral part of the lives of the human due to which it is more necessary to have a natural communication interface between humans and the computer. If we want to make the interaction between humans and computers more natural then we need to give the computers the ability to recognize the emotions of the humans through their speech.

Our main motive towards choosing this project and then making it is to understand underlying emotion through the way the humans deliver the speech. If the computer can able to understand the underlying emotion through speech then it would be easier for us to solve many problems related to depression, anxiety, etc. This also helps our intelligence agencies to put out the truth from the way of talking. Really, the computer has changed our living all throughout.

2. LITERATURE SURVEY

With Sentiment Analysis, we aim to identify and classify a voice or sound using information retrieval and computational linguistics. The opinion expressed is more significant than the topic on which the discussion is held.

When it comes to online shopping or any shopping in general, many among us consider the product reviews or ask people about their opinion on a product before making a decision. If other people's opinion changes our buying decision, it would definitely affect the business models of companies. This makes sentiment analysis very important. The tone of the customer's voice and expressions can determine the tone (positive, negative, or neutral) of their opinion. Sentiment Analysis categorizes this voice based on algorithms. It helps organizations to know how well they are doing in the market and also to work on their weaker areas in their business model that need attention.

Sentiment Analysis is the computational treatment of opinions, sentiments and subjectivity of the voice and has applications in different domains like business, politics, sociology and so on.

3. DATA SET

For the purpose of project 4 datasets were used.

3.1 CREMA-D

CREMA-D dataset is a collection of 7,442 original clips of 91 actors, 48 males and 43 females. Age group of 20 to 74. The dataset is associated with 12 sentences with

6 emotions - anger, disgust, sad, happy, neutral, fear

3.2 Ryerson audio-visual database of emotional speech and song

The dataset contains 1440 files: 60 trails per actor *24 actors, 12 males and 12 females. Speech emotions include calm, happy, sad, angry, fearful, surprise, disgust expression. Each expression is produced at two levels of emotional intensity (normal and strong) with add on neutral expression

3.3 Surrey audio-visual expressed emotion The SAVEE dataset was a recording of 4 male speakers who are post graduate and researchers at university of Surrey aged from 27 to 31 years. The dataset contains anger, disgust, fear, happiness, sadness and surprise emotion.

3.4 Toronto emotional speech set

There are a set of 200 targets. It consists of 2 actresses recording (aged 26 and 64 years) and consist of emotions such as anger, disgust, fear, happiness, pleasant surprise, sadness and neutral. There are 2800 data points in total.

4. MODEL

4.1 FEATURE EXTRACTION

Feature extraction is one of the most important part of analyzing and finding relations between different things. The audio file is a 3-D signal with three axes as frequency, time and amplitude.

1) Zero Crossing Rate: The rate sign changes the signal during the duration of a particular frame.

2) Chrome vector: A 12 element representation of the spectral energy.

3) MFCC: Mel frequency Cepstral Coefficient from a spectral representation where frequency band is not linear.

4.2 DATA AUGMENTATION

Data augmentation is a process by which we build new synthetic data sample by adding some perturbations on our initial training set. To generate augmented data for audio dataset we apply noise injection, pitch changes, stretch and changing time.

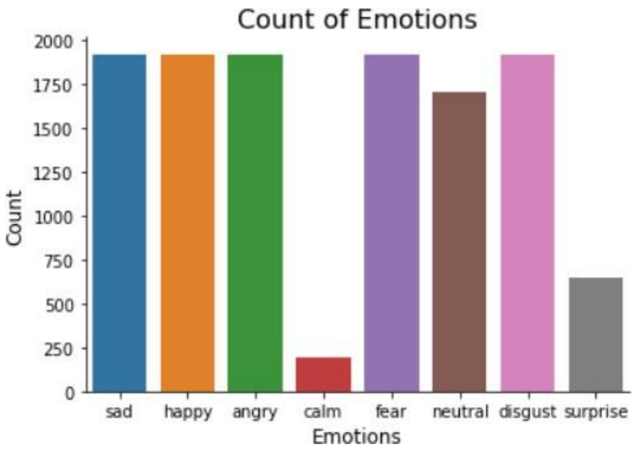


Fig -1: Emotion Count in Dataset

We can also plot wave plot and spectrogram. A wave plot is used to study the loudness of the audio signal. Spectrogram is a visual representation of frequencies of sound.

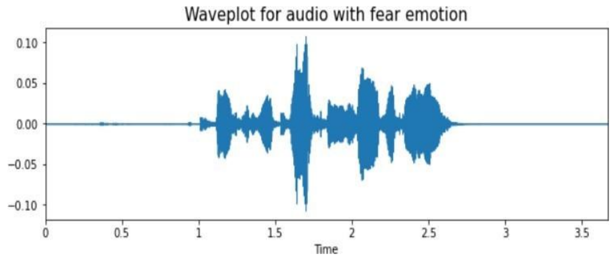


Fig -2: Wave plot for audio with fear emotion

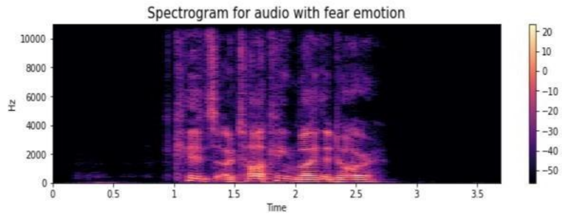


Fig -3: Spectrogram for audio with fear emotion

AUDIO DATA AUGMENTATION

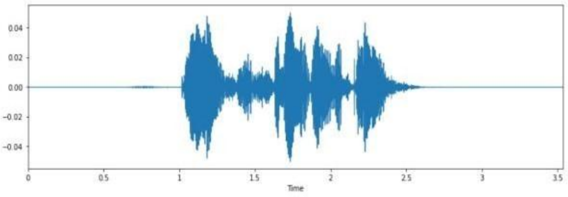


Fig -4: Wave plot for Simple Audio

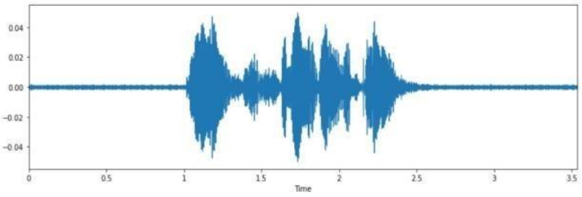


Fig -5: Wave plot for Noise Injection

4.3 MODEL BUILDING

For developing the model, we implemented CNN sequential model. Sequential model is a linear stack of layers. Sequential model is made by passing a list of layer instances to the model. It needs to be imported from keras.

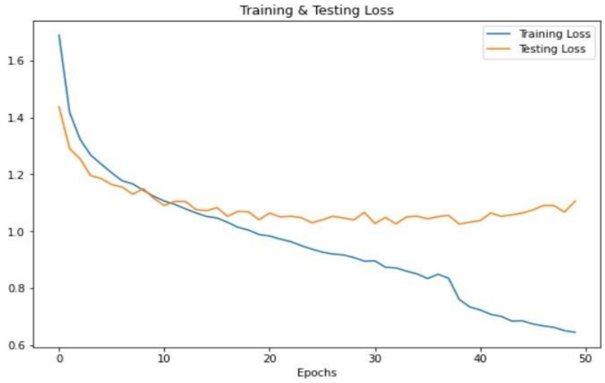
1) Conv1D - It means that the kernel can move only in one dimension along the axis of time.

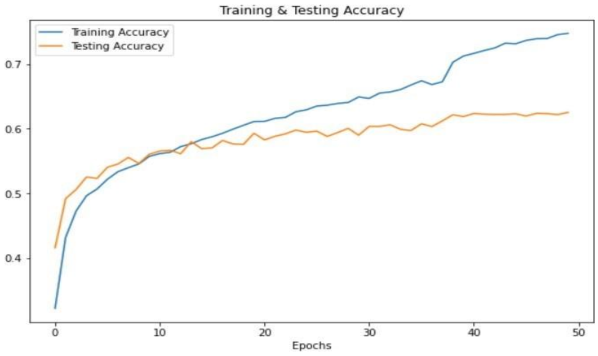
2) Kernel size- Kernel size refers to the width*height of the filter mask.

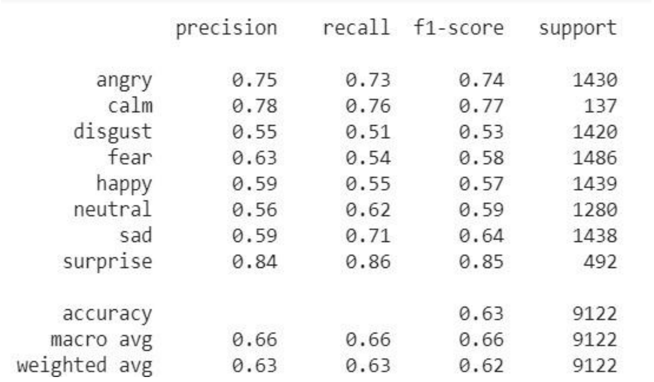
3) Padding - Padding refers to the quantity of pixels added to an image when it is being processed by the kernel of CNN.

4) Activation - Activation function is a node that is being added in the last or in between neural networks it is being added to decide whether the neuron would fire or not.

5) Dropout – Drop out layer is used to prevent the model from overfitting.







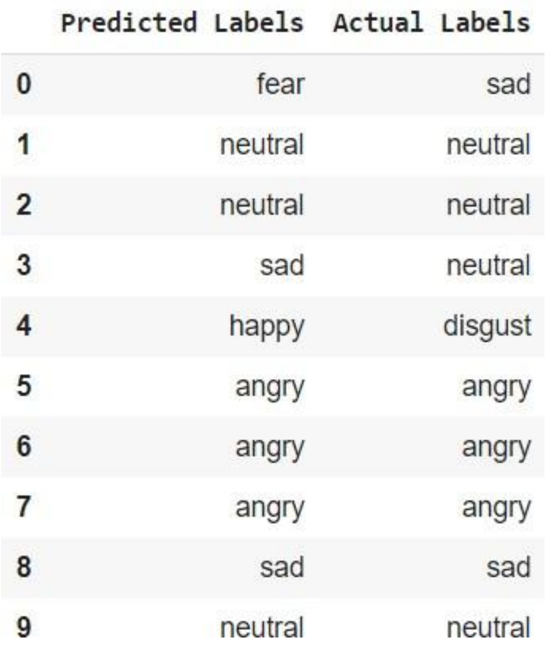
```
              precision    recall  f1-score   support

       angry       0.75      0.73      0.74      1430
        calm       0.78      0.76      0.77       137
     disgust       0.55      0.51      0.53      1420
        fear       0.63      0.54      0.58      1486
       happy       0.59      0.55      0.57      1439
     neutral       0.56      0.62      0.59      1280
         sad       0.59      0.71      0.64      1438
    surprise       0.84      0.86      0.85       492

    accuracy                           0.63      9122
   macro avg       0.66      0.66      0.66      9122
weighted avg       0.63      0.63      0.62      9122
```

CONCLUSION

Fig -6: Graphical Representation of DATA



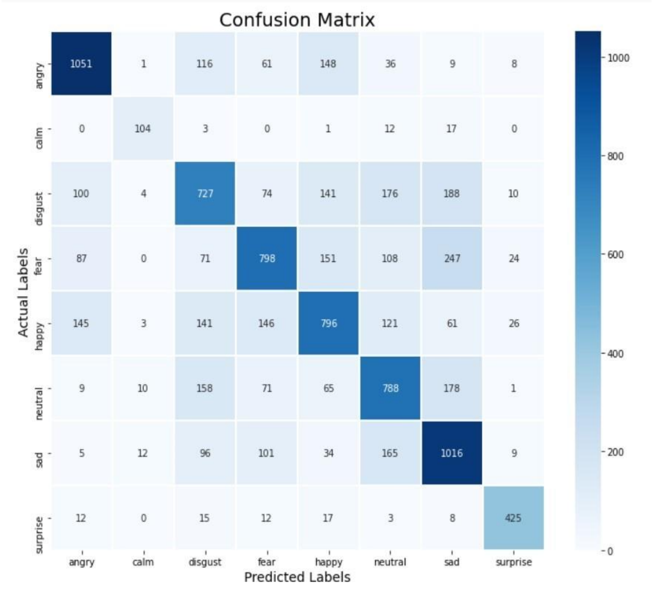
	Predicted Labels	Actual Labels
0	fear	sad
1	neutral	neutral
2	neutral	neutral
3	sad	neutral
4	happy	disgust
5	angry	angry
6	angry	angry
7	angry	angry
8	sad	sad
9	neutral	neutral

Companies can have a lot of valuable data that is unorganized data like emails, chats, social media, surveys, articles and documents. Going through all this data one by one is difficult, time-consuming and expensive too. Using sentiment analysis enables automation which makes it easier to gain valuable insights and change a particular business model accordingly. It processes vast amounts of information efficiently and at low-cost. One can also identify and alleviate a potential crisis.

Using machine learning algorithms, we can train a model to learn from the past data, so that it can predict an output for new data. The model becomes more accurate with more data.

FUTURE SCOPE

The accuracy of the model is not so high is due to limited application of augmentation. In further projects, we will train the data with more number of data augmentation processes.



REFERENCES

Number footnotes separately in superscripts. Place the actual footnote at the bottom of the column in which it was cited. Do not put footnotes in the abstract or reference list.

Use letters for table footnotes.

Unless there are six authors or more, give all authors' names; do not use "et al.". Papers that have not been published, even if they have been submitted for publication, should be cited as "unpublished" [4]. Papers that have been accepted for publication should be cited as "in press" [5]. Capitalize only the first word in a paper title, except for proper nouns and element symbols.

For papers published in translation journals, please give the English citation first, followed by the original foreign language citation [6].

1. G. Eason, B. Noble, and I. N. Sneddon, "On certain integrals of Lipschitz-Hankel type involving products of Bessel functions," Phil. Trans. Roy. Soc. London, vol. A247, pp. 529–551, April 1955. *(references)* J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68–73 S. Jacobs and C. P. Bean, "Fine particles, thin films and exchange anisotropy," in Magnetism, vol. III, G. T. Rado and H. Suhl, Eds. New York: Academic, 1963, pp. 271-350. K. Elissa, "Title of paper if known," unpublished.
2. R. Nicole, "Title of paper with only first word capitalized," J. Name Stand. Abbrev., in press.
3. Y. Yorozu, M. Hirano, K. Oka, and Y. Tagawa, "Electron spectroscopy studies on magneto-optical media and plastic substrate interface," IEEE Transl. J. Magn. Japan, vol. 2, pp. 740–741, August 1987 [Digests 9th Annual Conf. Magnetics Japan, p. 301, 1982].
4. M. Young, The Technical Writer's Handbook. Mill Valley, CA: University Science, 1989.