iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 10 **Issue:** VI **Month of publication:** June 2022

DOI: https://doi.org/10.22214/ijraset.2022.44583

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: 7.538
Volume 10 Issue VI June 2022- Available at www.ijraset.com

Speech Based Emotion Recognition

Ponnaluri Srinidhi[1], Mandala Yamini[2], Saikumar Bandi[3], Mahesh Kumar Punna[1]
[1, 2, 3]*IV B. Tech Students, Dept of IT, Sreenidhi Institute of Science and Technology(A), Hyderabad*
[4]*Assistant Professor, Dept of IT, Sreenidhi Institute of Science and Technology(A), Hyderabad.*

Abstract: In this paper we are exhibiting our Final Year Project which is Speech Emotion Recognition. Today's hot study issue is speech and emotion detection, with the goal of improving human-machine connection. Currently, the majority of research in this field relies on discriminator extraction to classify emotions into several categories. The majority of the present research focuses on the utterance of words that are employed in language-dependent lexical analysis for emotion detection. This study employs strategies to classify emotions into five categories: anger, calm, anxiety, happiness, and sorrow using Machine Learning algorithm Convolutional Neural Network.

I. INTRODUCTION

The approach of collecting emotion characteristics from computer voice signals, comparing them, and analysing the parameter values and the associated emotion changes is known as speech emotion recognition. To recognise emotions from audio sources, feature extraction and classifier training are necessary. The feature vector is made up of audio signal components that characterise the speaker's distinguishing characteristics (such as pitch, pitch, and energy) and is used to train the classifier model to properly detect certain moods.

In the social media and on the Web, there is a tremendous amount of opinionated data in the shape of Twitter, message boards, Facebook, blogs, and user forums.

Ideas shared on the internet by a varied collection of thought leaders and common individuals impact people's decision-making processes. Text-based reviews are one way for people to communicate their feelings/opinions about items or societal concerns. Another common approach to convey one's thoughts is through audio/video. Millions of videos on product and movie reviews, product unpacking, political and analysis of social issue, and opinion on these issues may be found on YouTube. There are several audio venues on the Internet where individuals may express themselves. In many cases, audio is more appealing than text because it gives more information about the speaker's opinions.

This vast resource is mostly underutilised, and collecting society sentiment/opinion on specific items, as well as mass opinion on social or political problems, will be incredibly useful for data analysis. Sentiment analysis in audio is still a relatively new area. Speech-based emotion extraction is a very young and challenging area. This study describes robust methods for extracting sentiment or opinion from natural audio sources.

II. RELATED WORK

Previous research used the highest cross correlation between audio recordings to categorise audio data into one of few emotion groups. As a result, more work in MATLAB has been done to determine the emotions of each audio recording supplied as an input. In MATLAB's toolbox, many classifiers are utilised, and classification learners categorise just a few emotion types.

Classification[1], Environment Sound-Classification[2], and Audio Generation[3] have all been accomplished with CNN-based models. Various 1-D convolution models have been created for working with raw audio waveforms; EnvNet[4] and Sample-CNN[1] are two examples.

However, CNNs on Spectrograms were used to derive the majority of the SOTA results. The majority of these models make the design more complicated by combining many models with diverse inputs and aggregating their outputs to produce predictions. [5] raw audio, spectrograms, and delta STFT coefficients were processed using three networks; [6] Are inputs to two networks, mel-spectrograms and MFCCs were used. We demonstrate however, that state of the art may be achieved using basic mel-spectrograms.

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 10 Issue VI June 2022- Available at www.ijraset.com

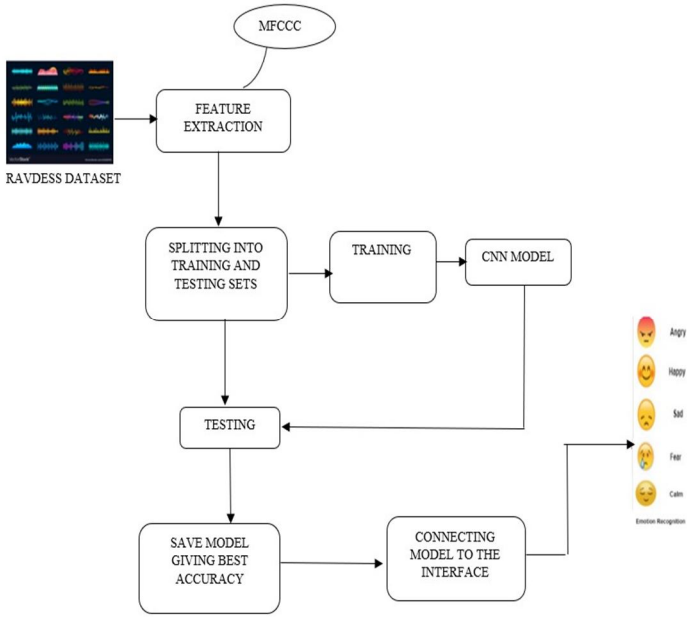
III. SYSTEM ARCHITECTURE

Models based on CNN have been employed in a variety of applications.

The System Architecture is a formal description and representation of system that is organised so that its structures and behaviours can be reasoned about. It shows the structure, behaviour, and other characteristics of a system. An architectural description is a formal description and representation of a system that is organised in such a way that the architecture and behaviours of the system are easier to comprehend. The components and subsystems of a system architecture are interrelated. This system is trained to recognize a person voice behaviour which consists of various emotions.

This Speech Based Emotion Recognition architecture flow consist of following

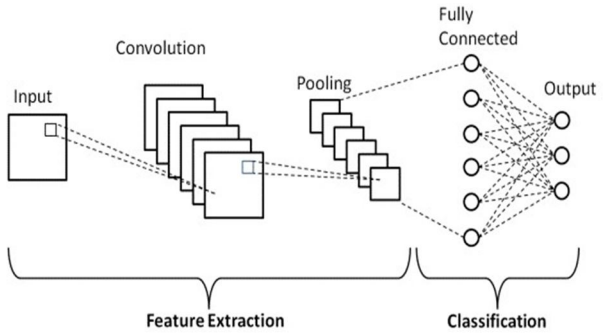
1) Feature Extraction
2) Splitting into Training and Testing sets
3) Testing
4) Save Model
5) Connecting model to the Interface
6) Emotion Recognition



IV. CNN MODEL

A CNN model is a learning system that can take a voice and prioritise (learnable weights and biases) particular aspects/objects while still identifying them. Other classification methods require a lot more pre-processing than a CNN. Although the hand-engineering of filters is required for basic operations, CNN may be able to learn these filters/characteristics with the correct training. Convolution is used to extract high-level input properties like edges. For CNN, a single Convolutional Layer is insufficient. The first ConvLayer is commonly used to capture low-level data like edges, colour, and gradient direction. With the addition of layers, the architecture responds to the attributes at the highest level, giving us a network that knows the photographs in the dataset as well as we do. Regardless of its flaws, convolutional neural networks have unquestionably ushered in a new era in artificial intelligence. Today, CNNs are used in face recognition, image search and editing, augmented reality, and other computer vision applications.

As advancements in convolutional neural networks develop, our findings are stunning and valuable. networks demonstrate, but we are still a long way from reproducing the core components of human intellect.



Convolutional neural networks stack artificial neurons. Artificial neurons, like actual neurons, compute the weighted sum of a vast number of inputs and output an activity value. A CNN's layers generate a large number of activation functions, which are passed on to the next layer.

1) As input, a sample audio file is provided.
2) To extract the MFCC, we utilise the LIBROSA python package (Mel Frequency Cepstral Coefficient)
3) Remixing the data, splitting it into train and test groups, and analysing the results training the dataset with a CNN model and its layers.
4) Predict human voice sentiment using the training data

The CNN model outperformed the others in the classification task. This trained model has the highest validation accuracy. There are a total of 18 levels.

The CNN model is divided into four layers:

a) *Convolutional Layer:* Detects significant regions at intervals, displays the feature map sequence, and recognises utterances of varied duration.
b) *Activation Layer:* As is common, a non-linear Activation layer function is applied outputs of the convolutional layer. The ReLu activation layer is used.
c) *Max Pooling Layer:* This layer provides the greatest possibilities for the Dense layers. It's simpler to keep variable-length inputs confined within a fixed-size feature array with this method.
d) *Flatten and Dense:* The pooled feature map is reduced to a single column, which is then transferred to the fully linked layer. Dense offers the neural network a layer that is totally coupled.

```
Model: "sequential"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
conv1d_7 (Conv1D)            (None, 216, 128)          768
_________________________________________________________________
activation_8 (Activation)    (None, 216, 128)          0
_________________________________________________________________
conv1d_8 (Conv1D)            (None, 216, 128)          82048
_________________________________________________________________
activation_9 (Activation)    (None, 216, 128)          0
_________________________________________________________________
dropout_3 (Dropout)          (None, 216, 128)          0
_________________________________________________________________
max_pooling1d_2 (MaxPooling1 (None, 27, 128)           0
_________________________________________________________________
conv1d_9 (Conv1D)            (None, 27, 128)           82048
_________________________________________________________________
activation_10 (Activation)   (None, 27, 128)           0
_________________________________________________________________
conv1d_10 (Conv1D)           (None, 27, 128)           82048
_________________________________________________________________
activation_11 (Activation)   (None, 27, 128)           0
_________________________________________________________________
conv1d_11 (Conv1D)           (None, 27, 128)           82048
_________________________________________________________________
activation_12 (Activation)   (None, 27, 128)           0
_________________________________________________________________
dropout_4 (Dropout)          (None, 27, 128)           0
_________________________________________________________________
conv1d_12 (Conv1D)           (None, 27, 128)           82048
_________________________________________________________________
activation_13 (Activation)   (None, 27, 128)           0
_________________________________________________________________
flatten_2 (Flatten)          (None, 3456)              0
_________________________________________________________________
dense_2 (Dense)              (None, 10)                34570
_________________________________________________________________
activation_14 (Activation)   (None, 10)                0
=================================================================
Total params: 445,578
Trainable params: 445,578
Non-trainable params: 0
_________________________________________________________________
None
The F1 Score is: 91.04
```

V. CONFUSION MATRIX

It's a table that's used to figure out where model flaws originated in classification problems. The rows reflect the category in which the results should have been. The columns, on the other hand, show our forecasts. It's simple to identify whose forecasts were incorrect with this table. We may utilise the confusion matrix function on our actual and expected values after the measurements have been imported.

Confusion_matrix = metrics.confusion_matrix
(actual, predicted).

To make the table more clear, we should transform it into a confusion matrix presentation.
cm_display = metrics.ConfusionMatrixDispl-
lay(confusion_matrix=confusion_matrix,
display_labels=[False, True])

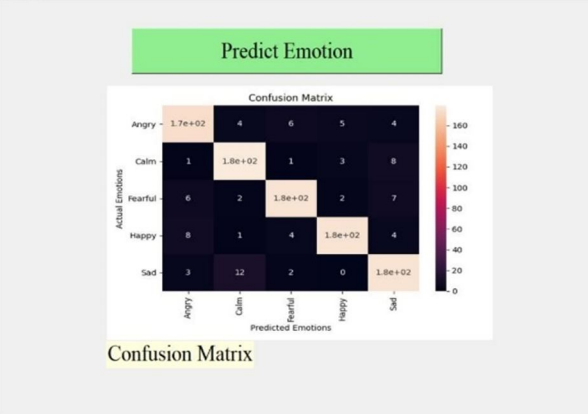


Fig 7.2 Recording & analysis of inputs

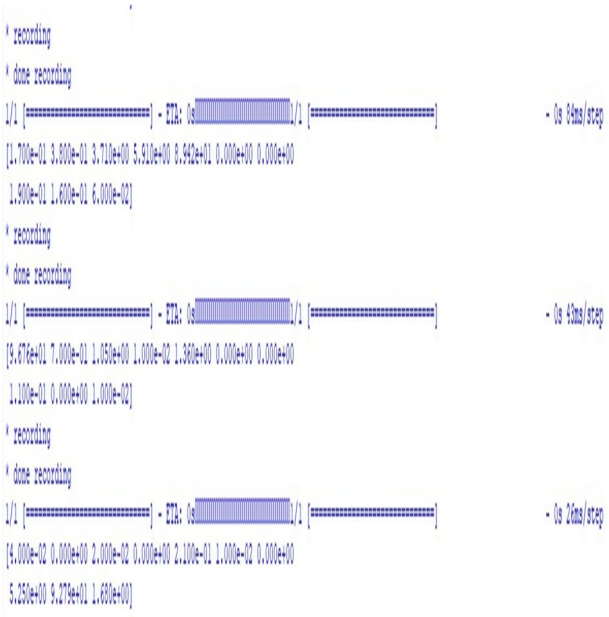


Fig 7.1 Prediction Page

VI. RESULTS

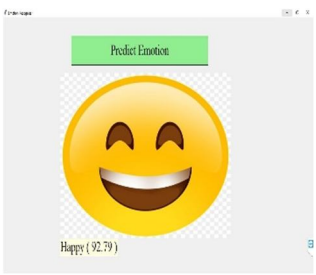


Fig. Happy Emotion

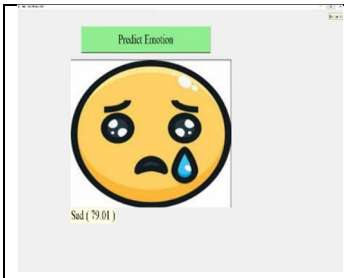


Fig. Sad Emotion

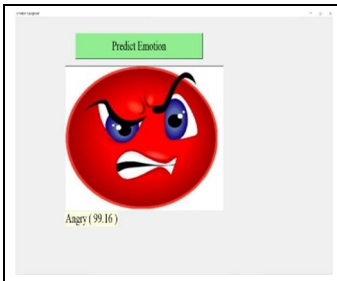


Fig. Angry Emotion

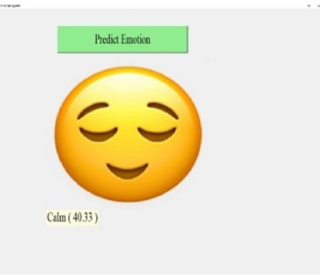


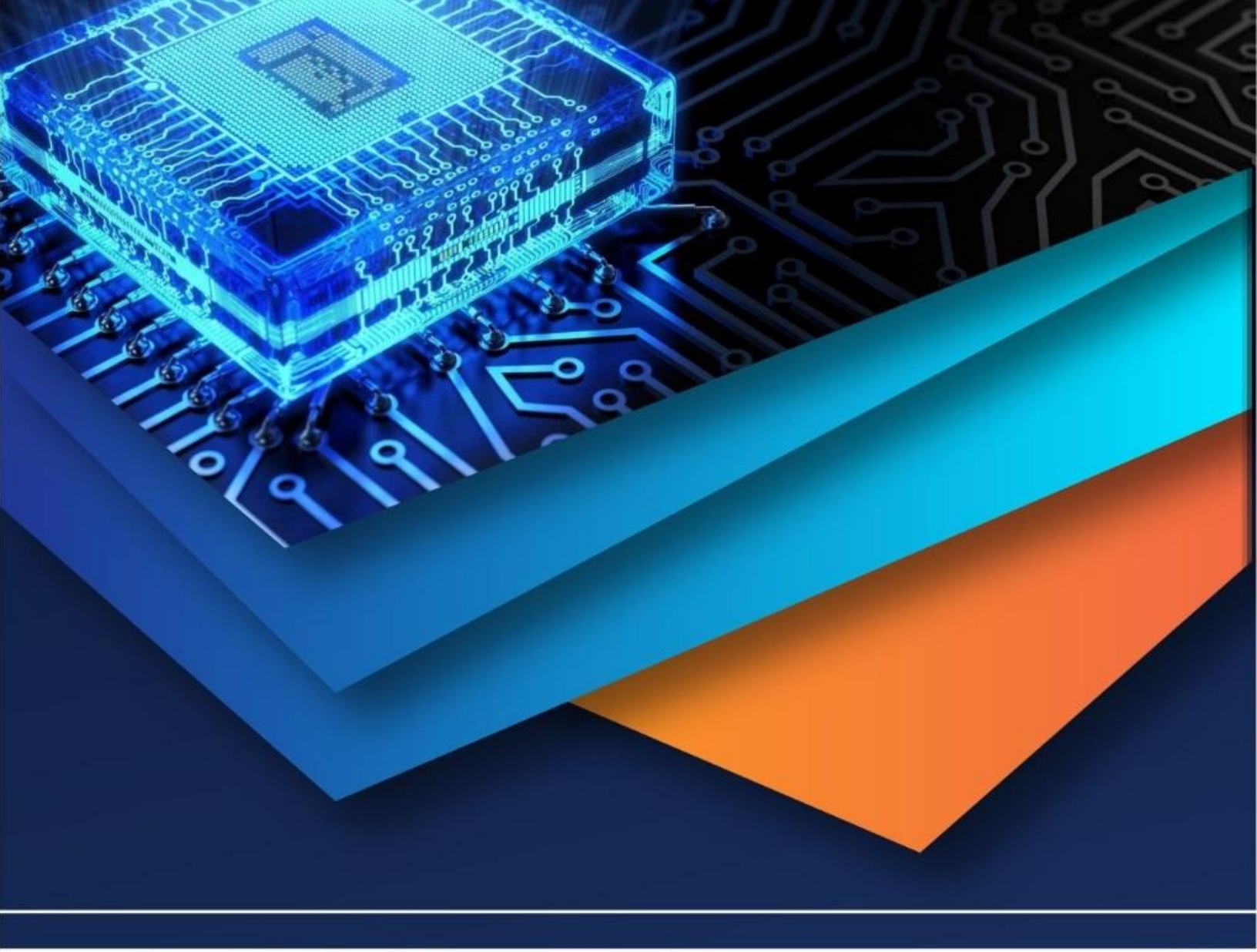
Fig. Calm Emotion

VII. CONCLUSION

This speech-based emotion recognition may be used to interpret the opinions/thoughts by feeding the audio into the model. For example, the sentiments they transmit about a product or a political viewpoint. This approach might be used in conjunction with a number of music apps to provide users with song recommendations depending on their emotions. This may also be used to enhance product suggestions for customers of online shopping applications like Amazon.

REFERENCES

[1] M. Dong, "Convolutional neural network achieves human-level accuracy in music genre classification," 2018.
[2] A. Guzhov, F. Raue, J. Hees, and A. Dengel, "Esresnet: Environmental sound classification based on visual domain models," 2020.
[3] A. v. d. Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu, "Wavenet: A generative model for raw audio," arXiv preprint
[4] Y. Tokozume and T. Harada, "Learning environmental sounds with end-to-end convolutional neural network," in 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2017, pp. 2721–2725.
[5] X. Li, V. Chebiyyam, and K. Kirchhoff, "Multi-stream network with temporal attention for environmental sound classification," 2019.
[6] Y. Su, K. Zhang, J. Wang, and K. Madani, "Environment sound classification using a two-stream cnn based on decision-level fusion," Sensors, vol. 19, no. 7, p. 1733, 2019
[7] A Research of Speech Emotion Recognition Based on Deep Belief Network and SVM Chenchen Huang, Wei Gong, Wenlong Fu, and Dongyu Feng https://www.hindawi.com/journals/mpe/2014/749604/
[8] https://vesitaigyan.ves.ac. in/recogni zing-emotion-from-speech-using-machinehttps://vesitaigyan.ves.ac.in/recognizing-emotion-from-speech-using-machine-learning-and-deep-learning/learning-and-deep-learning/



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY