



# IJRASET

International Journal For Research in  
Applied Science and Engineering Technology



---

# INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

---

**Volume:** 11    **Issue:** III    **Month of publication:** March 2023

**DOI:** <https://doi.org/10.22214/ijraset.2023.49448>

[www.ijraset.com](http://www.ijraset.com)

Call:  08813907089

E-mail ID: [ijraset@gmail.com](mailto:ijraset@gmail.com)

# Speech Based Emotion Recognition

C. Karthik Reddy<sup>1</sup>, T. Venkat Teja<sup>2</sup>, Imtiyaz<sup>3</sup>, Dr. Navnath D. Kale<sup>4</sup>

Dept of CSE, Vardhaman College of Engineering, Kacharam, Hyderabad

**Abstract:** *Speech Emotion Recognition is the final year project that we are showcasing in this essay. Speech and emotion recognition is a current research hot topic with the aim of enhancing human-machine interaction. In order to categorise emotions into different groups, discriminator extraction is now used in the bulk of this field of study. The majority of the current study concentrates on the words used in lexical analysis that is language-dependent for detecting emotions when they are spoken. This study uses the Convolutional Neural Network machine learning method five categories to classify emotions groups: anger, calm, anxiety, pleasure, and sorrow.*

## I. INTRODUCTION

Speech emotion identification is a method for identifying emotions by gathering their traits from computerized audio signals, contrasting them, and then examining the model parameters and associated emotional changes. Extracted features and classifier training are required in order to recognize and understand emotions from audio sources. The classifier model is trained to correctly identify specific moods using the feature set, which is made up of elements of the audio signal that describe the distinctive qualities of the speaker (such as pitch, pitch, and energy). There is a huge amount of opinionated content on social media and online, including on internet forums, Fb, blogs, and user forums. People's decision-making processes are influenced by the ideas posted on the internet by a diverse group of thought leaders and regular people. One way for consumers to express their thoughts and opinions about products or social issues is through text-based reviews. Audio and video are two more methods that are frequently used to communicate ideas. On YouTube, you may find millions of videos covering a wide range of topics, including products and film reviews, products unpacking, political and social issue analyses, and opinions on these topics. On the Internet, there are multiple audio platforms where people may express themselves. Audio frequently appeals to people more than text because that provides more insight into the speaker's viewpoints. Obtaining public attitude and opinion on certain issues as well as the general public's perception of social or political issues would be extremely helpful for data analysis. This enormous resource is typically underutilised. Audio sentiment analysis is still a relatively new field. The field of speech-based emotion extraction is still relatively new and difficult. In this work, reliable techniques are presented for extracting emotion or opinion from real-world audio sources.

## II. RELATED WORK

Earlier studies assigned audio data to one of a select few emotion categories based on the audio recordings with the strongest cross correlation. In order to ascertain the emotions of each audio recording provided as an input, extra work has been done in MATLAB. Just a small number of emotion types are classified by classification learners, who use a variety of classifiers from MATLAB's toolbox. There are several uses for CNN-based models, including categorization of audio and the surroundings [1, 2]. Just a few 1-D convolution models have been developed for using with raw audio waveforms, including EnvNet [4] and Sample-CNN [1]. Yet, the most of the SOTA discoveries were produced using CNNs using Spectrograms. Most of these models increase the difficulty of the design by using a number of models with varied inputs, whose combined outputs yield the predictions. As an example, [5] processed the raw sound, the spectra show, and even the delta STFT parameters using three networks; [6] utilised two networks only with input of a mel-spectrograms and even the MFCCs. We show, however, that conventional mel-spectrograms may be used to achieve cutting-edge performance.

## III. SYSTEM ARCHITECTURE

The System Architecture is an explanation and illustration of a system, set up in a way that facilitates discussion of the elements of the system's structure and behaviour. The architecture, behaviour, and other characteristics of a system are shown. A formal description of an architectural design expression and illustration of an established system such a way that it is easier to consider the structures and behaviours of the system. System parts and subsystems come together to form an architecture of the system and function as one cohesive unit. This system has been taught to distinguish a person's vocal behaviour, which includes a variety of emotions.

The steps in this architectural flow for speech-based emotion recognition are as follows:

- 1) Extraction of Features
- 2) Creating separate training and test sets
- 3) Testing
- 4) Preserving the model
- 5) Adding a model to an interface
- 6) Recognition of Emotions

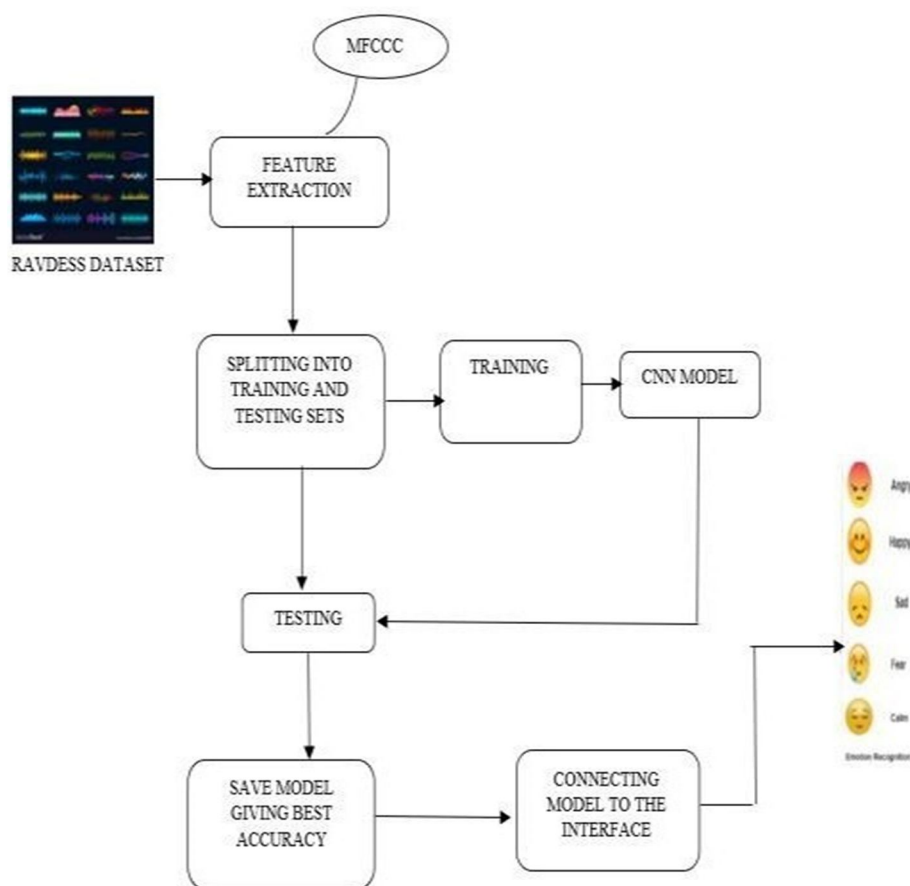


Fig. 1. Architecture of the system

#### IV. CNN MODEL

Convolutional Neural Network (CNN) models are learning systems that can take an input speech and recognise numerous features and objects while also giving priorities (learning weights and biases) to them. A CNN requires a substantially higher amount of data comparing to other classification techniques. smaller amount of preprocessing. With the proper training, CNN may learn these filters and properties, but for essential approaches, filters must be manually constructed.

Convolution operation is used to remove high-level input characteristics, such as edges. One Convolutional Layer is not sufficient for CNN. Low-Level data, such as edge, coloring, gradients direction, and so on, are typically collected by the first ConvLayer. The design responds to the highest-level attributes as well as the addition of layer, providing us a network that is just as familiar with the dataset's photos as we are. Notwithstanding their shortcomings, convolutional neural networks have without a doubt marked the beginning of an era in artificial intelligence. The usage of CNNs in computer vision applications such as augmented reality, face recognition, picture search, and editing, is widespread nowadays.

Our results are amazing and important for convolutional neural networks development. Networks show this, but we have a long way to go before we can replicate the fundamentals of human intelligence.

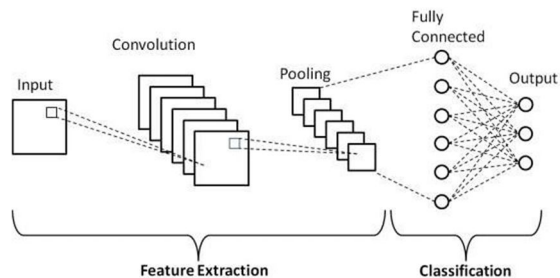


Fig. 2. CNN's Model layers

The layer of artificial neurons that make up the convolutional neural network are many. Artificial neurons are mathematical operations that calculate the weighted total of several inputs, then output an activation value, just like their biological counterparts.

The activation functions produced by a CNN's layers are sent to the following layer in a huge quantity.

- 1) The input consists of a sample audio file
- 2) The MFCC (Mel Frequency Cepstral Coefficient) is extracted using the LIBROSA Python library
- 3) Employing a Cnn architecture and its layers to train the dataset, reusing the data, splitting it into testing and training groups, and assessing the results
- 4) Use training data to predict the emotion of human speech In the classification challenge, the CNN model fared better than the competition. The most accurate validation is achieved by this trained model. There are 18 stages altogether. In the classification challenge, the CNN model fared better than the competition. The most accurate validation is achieved by this trained model. There are 18 stages altogether.

The CNN model is separated into four layers:

- a) Convolutional layer: recognises utterances of various durations, finds important sections at regular intervals, and shows the feature map sequence.
- b) Activation layer: As is typical for convolutional layer outputs, It uses a non-linear activating layer function. It makes use of the ReLu activation layer.
- c) Max Pooling Layer: For the Dense layers, this layer has the most potential. This approach makes it easier to maintain variable-length inputs inside a feature array with a fixed size.
- d) Flatten and Dense: A single column from the pooling layer map is then moved to the completely connected layer. Dense provides a fully linked layer to the neural network.

```

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 (MaxPooling1D) (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
    
```

Fig. 3. CNN Model four layers





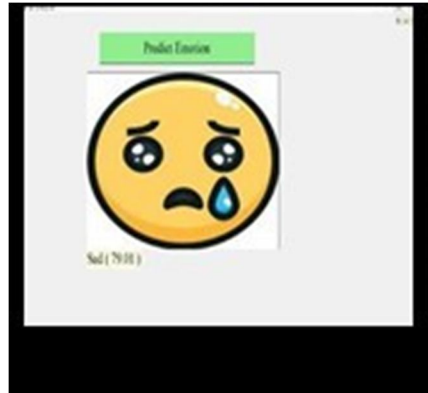


Fig. 7. Sad Emotion

## VII. CONCLUSION

By putting the audio into the model, this speech-based emotion identification might be utilised to decipher the opinions and ideas. Using their feelings towards a certain brand or political stance as an example. To provide consumers Providing listeners with music suggestions based on their feelings, this tactic utilised in conjunction with a variety of music applications. Moreover, this might be applied to improve the product recommendations sent to clients of online retailers like Amazon.



Fig. 8. Angry Emotion



Fig. 9. Calm Emotion

speech-using-machine learning-and-deep-learning/learning-anddeep-learning/



## 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/recognizing-emotion-from-speech-using-machine><https://vesitaigyan.ves.ac.in/recognizing-emotion-from->



10.22214/IJRASET



45.98



IMPACT FACTOR:  
7.129



IMPACT FACTOR:  
7.429



# INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Call : 08813907089  (24\*7 Support on Whatsapp)