INTERNATIONAL JOURNAL OF ADVANCE RESEARCH, IDEAS AND INNOVATIONS IN TECHNOLOGY

ISSN: 2454-132X
Impact Factor: 6.078
(Volume 7, Issue 3 - V7I3-1372)
Available online at: https://www.ijariit.com

Speech based Emotion Recognition using CNN Classifier

B. Sandeep
sandeepbavisetty111@gmail.com
Anil Neerukonda Institute of Technology and Sciences, Visakhapatnam, Andhra Pradesh

Dr. R. Sivaranjani
hod_cse@anits.edu.in
Anil Neerukonda Institute of Technology and Sciences, Visakhapatnam, Andhra Pradesh

R. Mourya
mourya.17.cse@anits.edu.in
Anil Neerukonda Institute of Technology and Sciences, Visakhapatnam, Andhra Pradesh

J. Sai Vinay
jsaivinay.17.cse@anits.edu.in
Anil Neerukonda Institute of Technology and Sciences, Visakhapatnam, Andhra Pradesh

Y. Vineela
yelisettyvineela.17.cse@anits.edu.in
Anil Neerukonda Institute of Technology and Sciences, Visakhapatnam, Andhra Pradesh

ABSTRACT

Communication through voice is one of the main components of affective computing in human-computer interaction. In this type of interaction, properly comprehending the meanings of the words or the linguistic category and recognizing the emotion included in the speech is essential for enhancing the performance. In order to model the emotional state, the speech waves are utilized, which bear signals standing for emotions such as boredom, fear, joy and sadness. This project is aiming to design and develop speech based emotional reaction (SER) prediction system, where different emotions are recognized by means of Convolutional Neural Network (CNN) classifiers. Spectral features extracted is Mel-Frequency Cepstral (MFCC). LIBROSA package in python language is used to develop proposed algorithm and its performance is tested on taking Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) samples to differentiate emotions such as happiness, surprise, anger, neutral state, sadness, fear etc. Feature selection (FS) was applied in order to seek for the most relevant feature subset. Results show that the maximum gain in performance is achieved by using CNN.

Keywords— *CNN, Audio Feature Extraction, LIBROSA, RAVDES, SER, MFCC*

1. PROBLEM STATEMENT

Emotions play a vital role in communication, the detection and analysis of the same is of vital importance in today's digital world of remote communication. Emotion detection is a challenging task, because emotions are subjective. We define a SER system as a collection of methodologies that process and classify speech signals to detect emotions embedded in them. Such a system can find use in a wide variety of application areas like interactive voice based-assistant or caller-agent conversation analysis. In this study we attempt to detect underlying emotions in recorded speech by analyzing the acoustic features of the audio data of recordings.

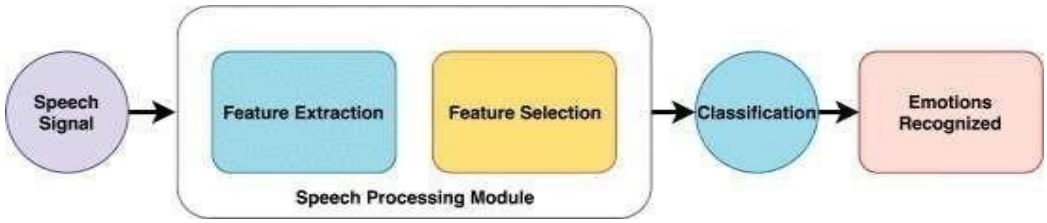
In this project, we will predict the emotion in the speech of a person's audio on the given dataset using CNN and deep learning algorithms. The dataset consists of 12,800 audio files of 12 male and 12 female voices with different emotions like happy, anger, sad, surprise, neutral, fear, disgust. The major goal of the proposed system is understanding Convolutional Neural Network, and predicting Emotion based on model.

2. INTRODUCTION

The human brain is an intricate organ that has been a lasting inspiration for research in Artificial Intelligence (AI). The neural networks in brain had the capability of learning all concepts from experiencing low level information and is remembers them which are processed by sensory periphery

The approach for speech emotion recognition (SER) primarily comprises two phases known as feature extraction and features classification phase. The first phase Feature extraction is the key part in the Speech Emotion Recognition. The quality of the features directly influences the accuracy of classification results. Typically, the Feature Extraction method designs handcraft features based on acoustic features of speech.

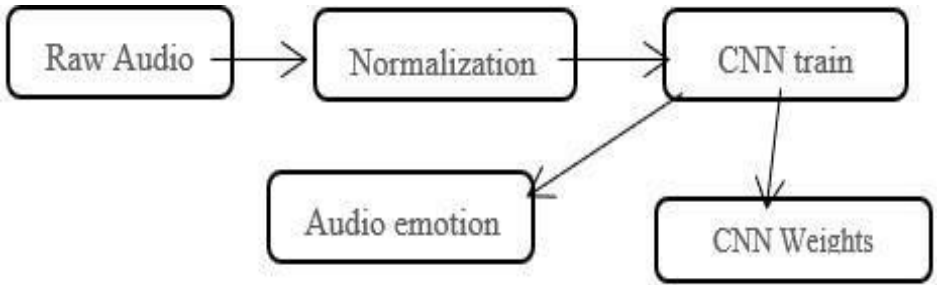
The second phase includes feature classification using linear and non-linear classifiers. The most commonly used linear classifiers for emotion recognition include the Maximum Likelihood Principle (MLP) and Support Vector Machine(SVM) and Convolution Neural Network (CNN). Usually, the speech signal is considered to be non-stationary. Hence, it is considered that non-linear classifiers.



3. METHODOLOGIES

3.1 Convolution Neural Network

The speech emotion recognition application is executed using CNN.Following is the architecture of the system:



3.2 Training Model and Testing Model

A training data is fetched to the system which consists the expression label and Weight training is also provided for that network. Anaudio is taken as an input. Thereafter, intensity normalization is applied over the audio. A normalized audio is used to train the Convolutional Network, this is done to ensure that the impact of presentation sequence ofthe examples doesn't affect the training performance. The collections of weights come out as an outcome to this training process and it acquires thebest results with this learning data. While testing, the dataset fetches the system with pitch and energy, and based on final network weights trained it gives the determined emotion. The output is represented in a numerical value and type of Emotion

Algorithm (Training):

1. Loading the data, dividing it in train and test
2. Using the LIBROSA, a python library we extract the MFCC (MelFrequency Cepstral Coefficient)
3. There after constructing a CNN model to train the dataset
4. Save Model

Testing and prediction

- The sample audio file(.wav) is provided as input.
- Using the LIBROSA, a python library we extract the MFCC(Mel Frequency Cepstral Coefficient)
- Predicting the human voice emotion from that trained data(predicted value)

Dataset

We are making use of RAVDESS dataset. It is downloaded from kaggle.com. It holds "12,800 files: 1800 audio-files/emotion multiplied with 7 different emotions = 12800 trials". The RAVDESS consists of 24 professional voices (12 feminine, 12 masculine), speaking2 lexically- matched sentences in the even North-American accent. Happy, sad, angry,fearful, calm, disgust and surprise are the various speech emotion expressions used. Every expression is generated in 2 levels of emotional intensity (light, bold), with a neutral expression. Every file out of 12800 files has an unique filename. The filename holds a 7-part numerical identifier (e.g., 03-01-05-01-01-01-01.wav)

Audio Feature Extraction

The Shape of the Speech signal determines what sound comes out. If the shape is determined accurately, then the correct representation of the sound being generated is obtained. The job of Mel Frequency Cepstral Coefficients' (MFCC's)is to correctly represent it. MFCCs is used as input feature. Loading and converting audio data into MFCCs format is done by python package librosa.

LIBROSA

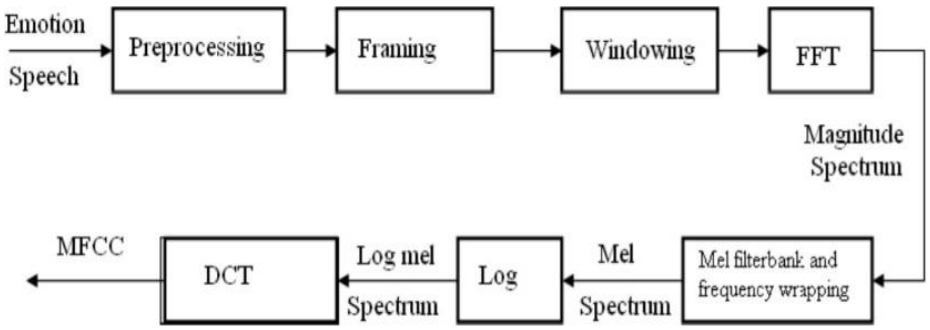
Librosa is a Python package for music and audio analysis. Librosa is basically used when we work with audio data like in music generation (using LSTM's), Automatic Speech Recognition. It provides the building blocks necessary to create the music information retrieval systems. Librosa helps to visualize the audio signals and also do the feature extractions in it using different signal processing techniques. It is the starting point towards working with audio data at scale for a wide range of applications such as detecting voice from a person to finding personal characteristics from an audio.

4. MEL-FREQUENCY CEPSTRUM

Mel-frequency cepstral coefficients (**MFCCs**) are coefficients that collectively make up an MFC. They are derived from a type of cepstral representation of the audio clip (a nonlinear "spectrum-of-a-spectrum"). The difference between the cepstrum and the mel-frequency cepstrum is that in the MFC, the frequency bands are equally spaced on the mel scale, which approximates the human auditory system's response more closely than the linearly-spaced frequency bands used in the normal spectrum. This frequency warping can allow for better representation of sound, for example, in audio compression. MFCCs are commonly used as features in speech recognition systems, such as the systems which can automatically recognize numbers spoken into a telephone. MFCCs are also increasingly finding uses in music information retrieval applications such as genre classification, audio similarity measures, etc

4.1 To train the model for accuracy calculation:

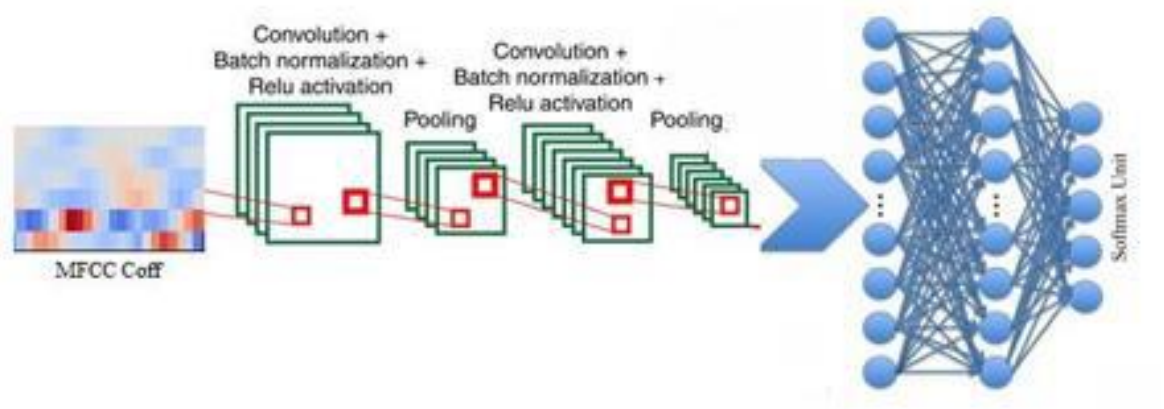
Within this module we train the model for accuracy estimations. First, import necessary modules and then import the dataset. We will receive the sampling rate value with librosa packages and MFCC function. Thereafter this value holds other variables. Now audio files and MFCC value hold a variable consequently it will add a list. Then zip the list and hold two variables x & y. Then we have represented (x, y) shape values with the use of numpy package.



Architecture of CNN

4.2 Implementation process of CNN model

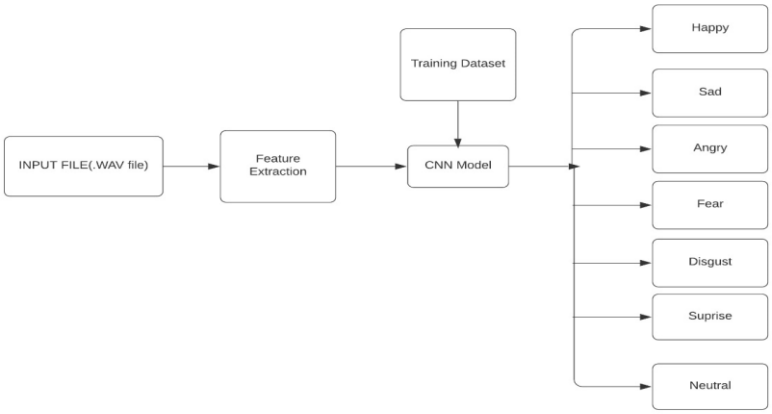
The deep neural network architecture actualized is convolutional neural network. In the proposed architecture after each convolutional layer max-pooling layer is placed. To establish non linearity in the model, for activation function Rectified Linear Units (ReLU) is used in both convolutional and fully connected layers. Batch normalization is used to improve the firmness of neural network, which normalizes the result of the preceding activation layer by reducing the number by what the hidden unit values move around and allows each of the layer in a network to learn by itself. Dense layer is used; in which all the neurons in a layer are connected to neurons in the next layer and it is a fully connected layer. SoftMax unit is used to compute probability distribution of the classes. The number of SoftMax to be used depends on number of classes to classify the emotions.



In our CNN model we have four important layers:

- Convolutional layer: Identifies salient regions at intervals, length utterances that are variable and depicts the feature map sequence.
- Activation layer: A non-linear Activation layer function is used as customary to the convolutional layer outputs. In this we have used corrected linear unit (ReLU) during our work.
- Max Pooling layer: This layer enables options with maximum value to the Dense layers. It helps to keep the variable length inputs to a fixed sized feature array.
- Dense layer: In Dense layer, for the perceptron we assign each node with input, weight and bias and we compute the function f(x)=activation (Weights * Input + bias). Before all this we perform ravel for data flattening. After computing the function for each node its output is given as input for the next layer or hidden layer and so on.. till the final output is received and then any of the activation functions are used and this output is compared to the real output and we find the error of the output. Now we apply back propagation to set the weights so that the error decreases resulting in good accuracy of the model. This is nothing but epoch.

4.3 Architecture of Speech Emotion Recognition



5. CONCLUSION

The model is trained with the training dataset and tested with the test data set, then the actual values are compared with the predicted values for accuracy of the model.

6. REFERENCES

[1] Automatic Speech Emotion Recognition Using Machine Learning By Leila Kerkeni, Youssef Serrestou, Mohamed Mbarki, Kosai Raoof, Mohamed Ali Mahjoub and Catherine Cleder, Published: March 25th 2019, DOI: 10.5772/intechopen.84856, https://www.intechopen.com/books/social-media-and-machine-learning/automatic-speech-emotion-recognition-using-machine-learning

[2] Speech Emotion Recognition Using CNN By Apoorv Singh, Kshitij Kumar Srivastava, Harini Murugan, DOI:10.37200/IJPR/V24I8/PR280260, https://www.researchgate.net/publication/342231090_Speech_Emotion_Recognition_Using_CNN

[3] Machine Learning Based Emotion Recognition using Speech Signal By K.Ashok Kumar, J.L.Mazher Iqbal, ISSN: 2249 – 8958, Volume-9 Issue-1S, Published 5, December, 2019, https://www.ijeat.org/wp-content/uploads/papers/v9i1s5/A10681291S52019.pdf

[4] SPEECH EMOTION RECOGNITION By Darshan K.A, U.B.D.T. College of Engineering, Davanagere, Karnataka, India and Dr. B.N. Veerappa, Department of Studies in Computer Science and Engineering, U.B.D.T. College of Engineering, Davanagere, Karnataka, India, Issue: 09, Sep 2020, https://www.irjet.net/archives/V7/i9/IRJET-V7I9154.pdf

[5] Audio-Textual Emotion Recognition Based on Improved Neural Networks By Linqin Cai, Yaxin Hu, Jiangong Dong and Sitong Zhou, Key Laboratory of Industrial Internet of ings & Networked Control, Ministry of Education, Chongqing University of Posts and Telecommunications, Chongqing, China, Published 31 December 2019, https://downloads.hindawi.com/journals/mpe/2019/2593036.pdf

[6] Speech Recognition using MFCC By Chadawan Ittichaichareon, Siwat Suksri and Thaweesak Yingthawornsuk, https://www.researchgate.net/publication/281446199_Speech_Recognition_using_MFCC?channel=doi&linkId=55e7e61d08aeb6516262ec4c&showFulltext=true

[7] For Dataset- Audio emotions, Sorted audio emotions from 4 data sets by Uldis Valainis https://www.kaggle.com/uldisvalainis/audio-emotions

