IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

SPEECH EMOTION DETECTION OF AUDIO USING DEEP LEARNING ALGORITHMS

[1] Prof. P.R. Patil, [2] Vishwajeet P. Salunke, [3] Suyash K. Lawand, [4] Shubham H. Talele, [5] Prathamesh S. Bandal

[1]Assistant Professor, Department of Computer Engineering, TSSM's Bhivarabai Sawant College of Engineering and Research, Narhe, Pune

[2] Department of Computer Engineering, TSSM's Bhivarabai Sawant College of Engineering and Research, Narhe, Pune

[3] Department of Computer Engineering, TSSM's Bhivarabai Sawant College of Engineering and Research, Narhe, Pune

[4] Department of Computer Engineering, TSSM's Bhivarabai Sawant College of Engineering and Research, Narhe, Pune

[5] Department of Computer Engineering, TSSM's Bhivarabai Sawant College of Engineering and Research, Narhe, Pune

Abstract: Analyzing sentiments using text has been present in the market since many days, but for processing audio data, we need to use deep learning algorithms. In case of audio files, we need to understand the audio features like the wavelength, frequency, pitch, etc. In this research, we have implemented the same and made use of these parameters to understand the audio emotions. The results are far better than text analysis and the model is capable of predicting audio emotions irrespective of language which has an advantage over traditional text-based Sentiment analysis**.**

Keywords **- Sentiment Analysis; Natural Language Processing; Feature Extraction; Machine Learning; CNN; Image Classification**

I. INTRODUCTION

In today's AI driven world, we see how close the AI has come to thinking and processing information just like humans do. Few years ago, AI was believed to be science fiction. But during the recent times AI is seeing a rapid growth in every sector and is performing better than expected [5].

The use of AI has increased significantly in many aspects like research, performing complex operations, etc. Among all other use cases it can be used to process and understand human behavior by understanding emotions in the conversations. Earlier, speech emotion detection included converting the audio into text in order to make predictions about the actual emotions. It worked well in many cases but it had a major flaw.

The models were not able to understand actual emotion behind the spoken word. For example, if someone said **"I am in an emergency, I need help!"** the person could be in panic situation but as the audio is converted to text it may not be able to properly interpret the information and it might make wrong predictions. Thus, to have a clear idea of what are the actual emotions behind the sentence in an audio we can use the features of sound like pitch, frequency, strength, etc. hence by having this information at hand we can easily make predictions based on these factors.

Opinion Mining or Sentiment Analysis is quite useful in monitoring of social sites as it gets a summary of the sentiments of the society on every topic. The ability to take out valuable information from social data is an exercise that is being extensively used by organizations all around the globe.

Some popular sentiment analysis applications include monitoring of social sites, management of customer support and analyzing customer response. Automatic sentiment analysis can be performed on any data source, to categorize survey responses and chats, Twitter and Facebook posts, or to scan emails and other documents.

All this is significant information for the companies and can make them take decisions accordingly. With its growing demand and advancement in sentiment analysis techniques, the analysis of sentiments is not only limited to textual data. Researchers are even exploring new possibilities in analysis of other modalities of data.

Today with the increased utilization of internet surfing, the enormous info produced is not only in textual form but more and more images and videos are being used to convey one's opinions. There has been significant amount of research on analyzing textual data but research related to other modalities of data including image, speech and video content has been limited

II. PREVIOUSLY AVAILABLE TECHNIQUES

There have been many ways through which we can make predictions for various emotions. They could be speech to text conversion for analysis, direct audio analysis, text analysis, etc. We will look into these methods in detail:

A. SPEECH-TO-TEXT CONVERSION AND ANALYSIS

In this method we mostly focus on the keywords for Detecting emotions, this method includes extracting the speech into text in the initial stage, The initial step involves text preprocessing, in which we remove any irrelevant information, such as special characters, punctuation or stop words [1].

Thereafter, we can tokenize the text by splitting it into individual words or phrases. The next step will involve extraction of the keywords and converting the text into numerical representation that can be used for analysis, which can be done with the help of various techniques such as **Bag-Of-Words** or **word embeddings**, or **TF-IDF.** Further steps involve utilizing these emotional dictionaries, which we have mapped earlier with the help of various techniques mentioned above. For analyzing the sentiments, we can use various deep learning algorithms, such as Naïve Bayes, Support Vector Machine or Neural Networks to classify the text into different emotional categories. And then finally, we can analyze and. Evaluate these predictions based upon various metrics such as F1 score, Recall, Analyzing confusion matrix etc.

B. EMOTION DETECTION USING AUDIO

In this technique, we analyze the emotions based on audio data, which involves the similar processes as like text analysis, such as audio preprocessing, feature extraction, creating classification Models using various algorithms like CNN, RNN, etc. [2]

1. *Preprocessing of audio files*.

The step involves gathering the audio files required for analysis and processing them to remove the irrelevant factors from audio such as noise, Irregularities, distorted audio, etc. This step also involves converting the raw audio file into the relevant format, such as WAV or FLAC to perform audio analysis.

2. *Audio Segmentation.*

Initially, the Raw audio files can be variable in length, which is a big issue in analyzing sentiments. Because, if the file is too big, it'll affect the quality of prediction, in case of larger audio files, the analysis becomes more difficult. And the results may vary as compared to expected results because of excessive diversity in the audio.

3. *Feature Extraction*

In this step, we extract the important features from the segmented audio clip Such as frequent Mel-Frequency Cepstral Coefficients (MFCC's), Pitch, Energy, Spectral Features, Duration, Wavelength. These features can now be used to analyze the Unique relations between various types of emotions and the audio features.

4. *Model Selection*

As the data is prepared by this step, we are now ready to perform the crucial step of model selection according to the requirement. There are various models present which are capable of emotion detection, such as CNN's, Support Vector Machines, Gaussian mixture models, random forests, Recurrent Neural Networks (RNN's), etc. We'll learn about the CNN model, as we will be using CNN model for this particular research topic.

a. Convolutional Neural Network (CNN)

CNN stands for Convolutional Neural Network, which is a type of deep neural network commonly used in computer vision tasks such as image recognition, object detection, and image classification. CNNs are inspired by the organization of the animal visual cortex and are designed to automatically and adaptively learn spatial hierarchies of features from input images. [5]

- **Convolutional Layers:** CNNs consist of multiple layers, including convolutional layers, which apply convolution operations to the input image. Convolutional layers use learnable filters (kernels) to extract features from the input image by sliding the filters across the image and performing element-wise multiplications and summations.

- **Pooling Layers:** Pooling layers are typically placed between convolutional layers. They down sample the feature maps obtained from convolutional layers by aggregating information within local regions. Common pooling operations include max pooling and average pooling, which retain the maximum or average value within each pooling region, respectively.

- **Activation Functions:** Activation functions such as ReLU (Rectified Linear Unit) are applied to the output of convolutional and pooling layers to introduce non-linearities into the network, enabling it to learn complex relationships in the data.

- **Fully Connected Layers:** Following the convolutional and pooling layers, CNNs often include one or more fully connected layers. These layers connect every neuron in one layer to every neuron in the next layer, enabling high-level feature representation and classification.

- **Training:** CNNs are typically trained using supervised learning, where the network learns to map input images to corresponding labels or categories. Training is achieved through backpropagation and optimization algorithms such as stochastic gradient descent (SGD) or its variants, which adjust the network's parameters (weights and biases) to minimize a loss function that measures the difference between predicted and true labels.

III. ACTUAL WORK

In this section, we will learn about the actual implementations which we carried out in this research.

1. LIBRARIES USED

a. Sklearn – for creating CNN model

b. Librosa – for audio pre-processing.

c. NumPy, Pandas – for mathematical operation.

d. TensorFlow – for machine learning models.

2. DATASETS

- **RAVDESS**

  This dataset contains 1440 files: 60 trials per actor x 24 actors = 1440. The RAVDESS contains 24 professional actors (12 female, 12 male), vocalizing two lexically-matched statements in a neutral North American accent.

  Speech emotions includes calm, happy, sad, angry, fearful, surprise, and disgust expressions. Each expression is produced at two levels of emotional intensity (normal, strong), with an additional neutral expression.

- **CREMA-D**

  It is a data set of 7,442 original clips from 91 actors. These clips were from 48 male and 43 female actors between the ages of 20 and 74 coming from a variety of races and ethnicities (African America, Asian,
  Caucasian, Hispanic, and Unspecified).

Actors spoke from a selection of 12 sentences. The sentences were presented using one of six different emotions (Anger, Disgust, Fear, Happy, Neutral, and Sad) and four different emotion levels (Low, Medium, High, and Unspecified)

- **TORONTO EMOTIONAL SPEECH SET (TESS)**

There are a set of 200 target words were spoken in the carrier phrase "Say the word _' by two actresses (aged

26 and 64 years) and recordings were made of the set portraying each of seven emotions (anger, disgust, fear, happiness, pleasant surprise, sadness, and neutral). There are 2800 data points (audio files) in total

3. DATA PREPERATION

In this phase we prepare the audio files from the dataset for further processing. This involves shifting of the audio, removing the noise in audio, and normalizing the factors of audio like pitch, wavelength, and frequency. This is done using Librosa library which provides us in-built functions for performing tasks such as **Time stretch, Pitch Shift, etc.**

4. FEATURE EXTRACTION.

In this step, we particularly extract the key features from the audio, such as Zero Crossing Rate ZCR, Short Time Fourier Transforms STFT, Mel-Frequencies. For this, we use Librosa Library, which allows us to extract these features using its inbuilt functions. In feature selection, we also perform dimensionality reduction, which involves reducing the features of audio which are irrelevant and computationally complex to perform. And hence, we use **Principal Component Analysis (PCA)** or other feature selection algorithms which can be used to reduce the dimensionality of the audio and select the important features.

5. MODEL SELECTION

- ***LSTM model:***

Long-Short Term Memory is a type of recurrent neural network architecture designed to overcome the limitations of traditional RNN's in capturing long term dependencies in sequential data.

LSTM is composed of cell state, input gate, forget gate and output gate. The cell state is nothing but a long-term memory

component that runs through entire chain of LSTM units. It is selectively updated and modified through gates, ensuring

relevant information persists while irrelevant information is discarded.

- **Forget Gate:** Forget gate determines what information to discard and. It takes the previous hidden state and the current input as input and outputs a forget vector.
- **Input Gate:** It consists of two parts, a sigmoid layer called input gate layer that decides which value to update, and a tan-h layer that creates a vector of new candidates to be added to self-state
- ***CNN model:***

A convolutional neural network is a type of deep neural network specifically. Designed for processing structured grid like data, such as image and audio CNNs are highly effective in task involving feature learning from raw data, however. They have also been successfully applied to other domains, including natural language processing and audio analysis. [5]

The CNN model includes convolutional layers, pooling layers, activation functions, fully connected layers and dropout functions. The convolutional layers operate on the input data using kernels and enable the network to extract spatial hierarchies of features from the input data. In this particular research paper, we have focused on CNN model as it has more flexibility from learning through raw data than LSTM model.

6. TRAINING

We have used the available datasets like TESS, RAVDESS, CREMA-D to train the model. Each of these datasets possess unique audio features which will help the model to explore more about the available emotions and it can have variety of data to explore and establish relations of emotions in the audio.

During the training we introduced noise, shift, change in frequency in the training audio data to make the model more adaptable to handling real world audio. As we know the audio might have certain noise, stretch, varying wavelengths and frequencies. And it is necessary to train the model to adapt these changes so that it won't make biased predictions on real data.

7. VALIDATION

For evaluating the trained model on the test dataset, we use calculation metrics such as accuracy, precision recall, F1 score to quantify the model's effectiveness. We will learn about these metrics in brief. [2][5]

- **Precision:** Precision measures the proportion of true positive predictions among all positive predictions made by the classifier

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives}+\text{False Positives}}$$

- **Recall:** Recall measures the proportion of true positive predictions among all actual positive instances in the dataset.

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives}+\text{False Negatives}}$$

- **F1 Score:** The F1 score is the harmonic mean of precision and recall.

$$F1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision}+\text{Recall}}$$

IV. RESULTS AND DISCUSSION

After studying multiple models like LSTM, Random Forests, ANN's and CNN. We discovered many changes in factors like performance, computational costs, accuracy, effectiveness and speed.

While studying about various ways through which we can analyze speech emotions, we computed the accuracy of the few available models like LSTM, Random Forests, ANN's and CNN's. The difference between accuracy can be observed in the given table 4.1

Table 4.1: Comparison between accuracy of different models

Sr. No	Model	Accuracy
1.	LSTM	78.59%
2.	CNN	92.65%
3.	ANN	83.41%
4.	Random Forest	81%

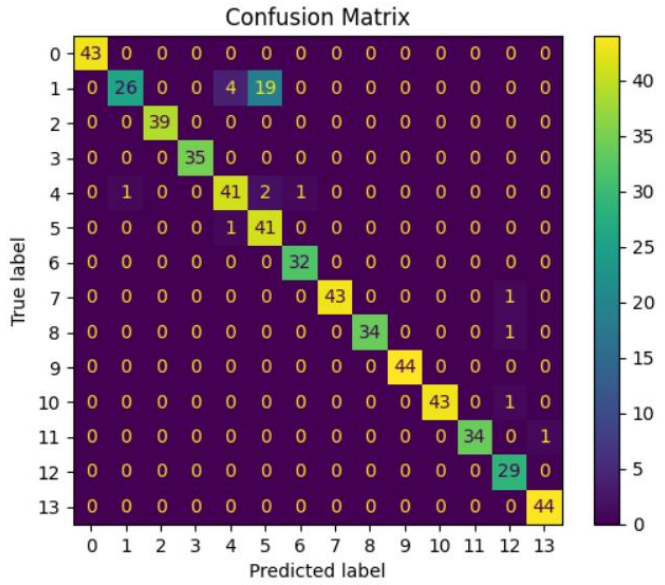

Fig 4.1 Confusion Matrix of CNN model after testing the model on available datasets

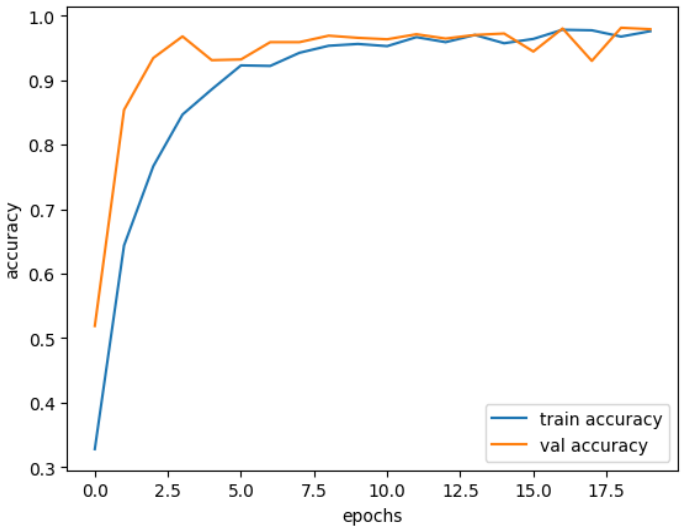

Fig 4.2 Model Accuracy Graph

We can analyze the performance and accuracy of a model after testing with help of confusion matrix and other parameters like F1 score, Precision, Recall, etc. the confusion matrix of CNN model can be seen in **Fig 4.1** where we can see the heatmap indicating the relation between actual results and model predictions.

The accuracy and loss while model training can be observed in **Fig 4.2**. as we can observe the model has been trained well as there is an exponential increase in the training accuracy

IV. CONCLUSION

During this research of sentiment analysis of audio using deep learning we came to know about various techniques used for sentiment analysis like audio to text conversion, direct audio analysis, etc. We learned each of them in brief and compared their performance based upon various metrics of validation, such as F1 score, precision, recall, etc. While we were learning about these techniques, we also came across different Deep Learning algorithms, such as Neural Networks, Random Forests, Support Vector Machines, etc. After learning about these different emotion analysis methods. Finally, we can conclude our research.

V. REFERENCES

Various sources which were used while conducting this study are listed below. These include research papers and journals from other authors and other available articles over the web.

[1] Research Paper on **"Sentiment Analysis of Text and Audio Data"** by Dr. Munish Mehta, Kanhav Gupta, Shubhangi Tiwari, Anamika

[2] A Research Paper on **"AUDIO SENTIMENT ANALYSIS"** by P.Ansar khan, T.sumanth, K.vishnu Vardhan

[3] A Research paper on **"Audio and Text Sentiment Analysis of Radio Broadcasts"** Naman Dhariwal, Sri Chander Akunuri, Shivama, And K. Sharmila Banu

[4] A paper on **"AUDIO SENTIMENT ANALYSIS"** by Dr. C.S. Shinde, Sumit Sadashiv Kapase, Saurabh Ashwinkumar Shetti, Sumedh Malappa Khurapi, Shubham Pandurang Desai, Asim Akabar Sayyad

[5] Web Source: Wikipedia, Google Search