International Journal of Engineering Applied Sciences and Technology, 2022
Vol. 6, Issue 11, ISSN No. 2455-2143, Pages 181-184
Published Online March 2022 in IJEAST (http://www.ijeast.com)

SPEECH EMOTION RECOGNITION USING MACHINE LEARNING AND DEEP LEARNING

Utkarsh Kumar Singh, Sudhanshu Singh, Shilpi Khanna, Radhey Shyam
Department of Computer Science and Engineering,
Shri Ramswaroop Memorial College of Engineering and Management, Lucknow, Uttar Pradesh

***Abstract*— This paper presents a study of human emotions emitted through sound or speech. Perceiving a person's emotions through sound has always been a difficult task for the machines. If a machine can recognize the emotion emitted by its user, it would be easier for it to help the user in a better way to perform the task he needs to do.**
Many researchers have worked on this problem so far. Some have classified the emotions as four basic human emotions: "Happy", "Sad, "Angry" and "Neutral". And there has been usage of dimensional aspects such as Valence (Positivity), Activation (Energy) and Dominance (Controlling impact) to detect emotion using speech.
Speech emption recognition is quite difficult for machine learning as the analysis of sound and speech signals is difficult to do as it includes a plethora of frequencies and features.
We have also done a comparative study on different modes and datasets namely, RAVDESS, IEMOCAP datasets.
We have implemented Multi-Layer-Perceptron model which is performing at 66.2% accuracy for the emotions such as happiness, sadness, anger, neutral, calm, disgust, etc.
Our evaluation shows that the proposed approach yields accuracies of 65.8%, 66.2%, and 63.2% using RF, MLP Classifier and SVM Classifiers, respectively.

***Keywords*— multimodal speech emotion recognition, machine learning, deep learning, emotions, random forest, multilayer perceptron, extreme gradient boosting, support vector machine**

I. INTRODUCTION

"Sound is a vibration that propagates as an acoustic wave, through a transmission medium such as a gas, liquid or solid. In human physiology and psychology, sound is the reception of such waves and their perception by the brain" [11].
"Emotions are biologically-based psychological states brought on by neurophysiological changes, variously associated with thoughts, feelings, behavioral responses, and a degree of pleasure or displeasure" [12].
We know when a person is happy, sad, angry, surprised or even neutral by simple look at his/her face if we are meeting in person, listening to his voice while on call or even by his picture which was taken years ago. Non-verbal sounds within an utterance also play an important role for us for recognizing emotions.
Machine learning is a rapidly growing field in computer science having state-of-the-art applications in several other fields of study and is being implemented commercially. Machine learning is a technique of training machines to perform the activities a human brain can do, although a lot faster and better than an average human-being is capable of [4]. Here, we are going to use machine learning models to train computers to detect human emotions by their speech.
The sound emotion database can be constructed by three major ways: Real Data (Natural speech by people), Mining data from social media (MSP Podcasts and crowdsourcing) or to ask people (actors) to act or speak with different emotions (IEMOCAP data).
Emotion Recognition Theory: "Methodology to describe the emotion so that we can label the data where is the appropriate target for machines to predict"
Human Emotion Recognition is a simple task if we are considering the recognition done by human mind. It is simple and easy. It is not taught to us by anyone but we keep learning about different emotions since we are a little child, subconsciously.
The speech emotion recognition problem as a substantial potential for utilization in many applied industries, being robotics, cell phone calling, customer satisfaction software and other systems with iterative user interaction.
If we solve this problem, it will allow us to receive user's feedback in a natural way without any supplementary user's actions. It will also help us to speed up the computer-person interaction.
"An experimental result on the recognition of seven emotional states in the NNIME (The NTHU-NTUA Chinese multimodal emotion corpus) showed that their proposed method achieved a detection accuracy of 52.00 % outperforming the old traditional methods" [1].
Other researchers have also reached the accuracy of about 70 to 80 percent in their researches.

II. METHODOLOGY

The first step comprises of gathering the emotional dataset from different sources. Extracting database from real time data such as natural conversations is difficult as we do not always transmit emotional sound, rather a major portion of the emotions transmitted by humans are non-verbal like gestures,

International Journal of Engineering Applied Sciences and Technology, 2022
Vol. 6, Issue 11, ISSN No. 2455-2143, Pages 181-184
Published Online March 2022 in IJEAST (http://www.ijeast.com)

motions, facial expressions, etc. So, to overcome this issue we can extract emotionally rich data from different social media platforms and from movies and podcasts. Other than this, we can get actors to enact different emotions in different voices which is the most commonly used method for emotional data collection.

After this, we have to extract the features from labelled data, select important and meaningful features based on various contexts, moods and sources for proper classification.

Following this, we will use popular machine learning algorithms to classify the data and detect the emotions of the speaker.

We will also build a User Interface for user to test the model in real time.

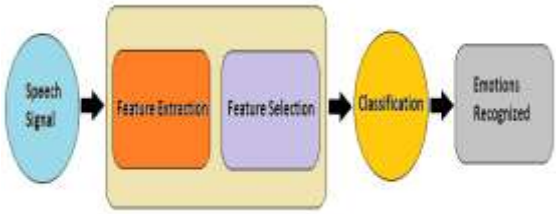

Fig. 1. Methodology

III. MACHINE LEARNING MODELS

A. **Random Forest Model**: Random Forest models make use of multiple decision trees while training and output a mode of classes (classification) of each tree. Classification may be defined as the grouping of the similar elements into a class. If the algorithm classified the labels of input into two separate classes, it is referred as binary classification [3]. It works on the following two principles:

1. Each Decision tree predicts using a random subset of features.[2]
2. Each of decision tree is trained with only a subset of training samples. [5]

B. Gradient Boost/ extreme Gradient Boosting: It is also a classifier using decision trees, trained in sequential manner, using forward stage-wise additive modelling. The decision trees learn simply in the early stages. As the training progresses, it becomes more powerful as it focuses on instances where previous learners made errors. At the end of training, the prediction is a weighted linear combination of the output from individual learners [6].

C. Support Vector Machine: SVM is a set of supervised learning method for classification, regression and outlier detection. In SVM, each data item is plotted in n-dimensional space with the value of each dimension being the value of a particular coordinate. Then, classification is performed and we find the hyper-plane and differentiate the two classes [14].

D. Logistic Regression: LR is typically used for binary classifications [7]. It helps us in understanding the relationship between a dependent variable and multiple independent variables using logistic regression equations.

E. Multinomial Naïve Bayes Model: MNB is a classifier suitable for classification with discrete features. It calculates the probability P(c/x) where c is the class of possible outcomes and x is the instance to be classified.

IV. DEEP LEARNING MODEL

A. Artificial Neural Network – Multilayer Perceptron (MLP) Model

Artificial Neural networks (ANN) are equipped fir learning any non-linear function. The building blocks of neural networks include neurons, weights and activation functions. These can be considered as a powerful technique in voice recognition [15].

Artificial neural networks can learn loads that map any contribution to the yield. These neural networks have the ability to learn through training and observation. Through training, the network builds up a relationship between the inputs and the outputs. Parameters like weights and biases are tuned which minimizes the error. An MLP is made of minimum of three layers namely, Input Layer, Hidden Layer and the Output Layer. The outputs of each layer are connected to the input of the next layer. These are majorly used for classification problems.

The strength of neural networks comes from their ability to study the representation in your training data and how to best relate it to the output variable which you need to predict. In this experience neural networks analyze a mapping. Mathematically, they're able to gaining knowledge of any mapping function and have been verified to be a universal approximation algorithm.

Each and every node of an MLP uses a particular type of activation function. The initial node uses a linear function whereas all the other nodes use non-linear functions. MLP is trained by a process called as “backpropagation”. It has the ability to distinguish different types of data. For complex tasks, multiple layers are used. The complexity of the model is directly proportional to the number of layers used in the MLP. Inputs are categorized using straight lines. The input is a vector which is multiplied by a load ‘w’ and added to a constant ‘b’ called bias. [8].

Mathematical Equation for the same is:

$$y = \varphi\left(\sum_{i=1}^{n} w_i x_i + b\right) = \varphi(w^T x + b) \qquad (1)$$

Let us get a clear picture of the Multi-Layer Perceptron model with a diagram.

International Journal of Engineering Applied Sciences and Technology, 2022
Vol. 6, Issue 11, ISSN No. 2455-2143, Pages 181-184
Published Online March 2022 in IJEAST (http://www.ijeast.com)

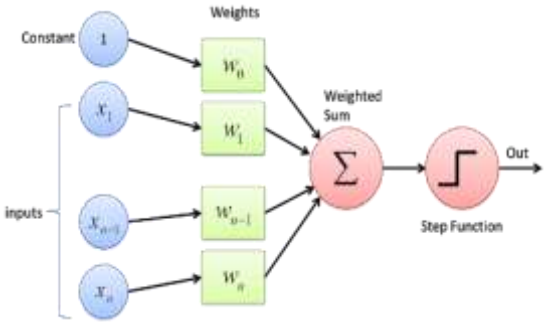

Fig. 2. Perceptron Structure [13]

V. IMPLEMENTATION

Here we will describe the details of the implementations in the work.

a) We used python's librosa [9] library to process audio and extract features.
b) Python's Scikit-learn and xgboost [10] libraries are used to implement the machine learning classifiers like Random Forest, Gradient Boost, SVM, Naïve Bayes Classifier and Logistic Regression).
c) We used Jupyter Notebook for our implementations.
d) We spilt the data randomly into 80% training data and 20% test data. We also collected data from our friends and colleagues to test the data in real time.
e) We developed a User Interface using Python's Django framework on through which a user can upload his voice and the UI will show the emotion as output.

VI. EXPERIMENT AND RESULT

We trained and tested all the above-mentioned Machine Learning and Deep Learning Models.

a) Audio-only: Here, we trained all models using the audio-only feature vectors.
b) Text-only: We trained all models using text-only features.
c) Audio + Text: We combined the feature vectors from the above two experiments. There are methods to combine the vectors efficiently from multiple modalities. a high impact on the model's performance.

The performance of the models is described in this section.

a) For Audio-Only

Table -1 Experiment Result

Model	Accuracy
RF	56.9
XGB	57.1
SVM	40.1
LR	32.3
MNB	30.6
MLP	40.1

These are the individual scores of the models. We also combined XGB, MLP and RF models together and got the accuracy of 57.6%.

b) For Text-Only

Table -2 Experiment Result

Model	Accuracy
RF	60.8
XGB	55.8
SVM	61.5
LR	62.1
MNB	59.6
MLP	61.0

The combined accuracy of XGB, MLP, RF, MNB and LR is 62.6

c) For Audio + Text

Table -3 Experiment Result

Model	Accuracy
RF	65.8
XGB	63.3
SVM	63.2
LR	63.5
MNB	60.7
MLP	66.2

The combined accuracy of XGB, MLP, RF, MNB and LR is 70.1%.

VII. USER INTERFACE

We used Python's Django framework to develop a User Interface to allow end users to test their audio and detect emotions. The user can upload the file in .wav format and he will get the output of the emotion he had in the audio.

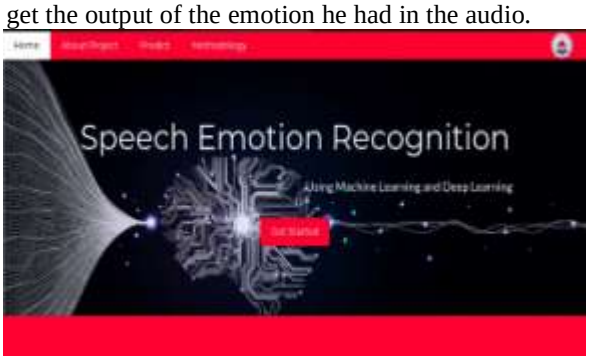

Fig. 3. User Interface Home Page

The user can click on "Get Started" button or "Predict" button on the navigation bar to be redirected to the prediction section. He can then upload his audio file in .wav format and click on "Predict Emotion" to get the output. We can see the working in the following screenshot of the user interface.

International Journal of Engineering Applied Sciences and Technology, 2022
Vol. 6, Issue 11, ISSN No. 2455-2143, Pages 181-184
Published Online March 2022 in IJEAST (http://www.ijeast.com)

Fig. 4. Predicted Emotion - Disgust

VIII. CONCLUSION

In this paper, we studied the speech emotion recognition by machines through various models on a variety of datasets. We also built a user-friendly UI which would help us to test the data easily without opening the code editor. We compared different Machine Learning and Deep Learning models. We also combined multiple models to obtain a higher accuracy than the individual models. Overall, it was a great experience for us working on this project and we would be interested in this project in future to add on more data that we collect eventually.

IX. REFERENCE

[1]. Kazheen, Ismael & Mohsin Abdulazeez, Adnan. (2021). Deep Learning Convolutional Neural Network for Speech Recognition: A Review. 10.5281/zenodo.4475361.

[2]. Y. Amit, D. Geman, and K. Wilder, "Joint induction of shape features and tree classifiers," IEEE Transactions on Pattern Analysis & Machine Intelligence, no. 11, pp. 1300–1305, 1997.

[3]. Radhey Shyam, Ria Singh. A Taxonomy of Machine Learning Techniques. Journal of Advancements in Robotics. 2021; 8(3): 18–25p.

[4]. Radhey Shyam, Riya Chakraborty. Machine Learning and its Dominant Paradigms. Journal of Advancements in Robotics. 2021; 8(2): 1–10p.

[5]. L. Breiman, "Bagging predictors," Machine learning, vol. 24, no. 2, pp. 123–140, 1996.

[6]. J. Friedman, T. Hastie, and R. Tibshirani, The elements of statistical learning, vol. 1. Springer series in statistics New York, 2001.

[7]. G. King and L. Zeng, "Logistic regression in rare events data," Political analysis, vol. 9, no. 2, pp. 137–163, 2001.

[8]. S. Sharma, "Emotion Recognition from Speech using Artificial Neural Networks and Recurrent Neural Networks," 2021 11th International Conference on Cloud Computing, Data Science & Engineering (Confluence), 2021, pp. 153-158, doi: 10.1109/Confluence51648.2021.9377192.

[9]. B. McFee, C. Raffel, D. Liang, D. P. Ellis, M. McVicar,E. Battenberg, and O. Nieto, "librosa: Audio and music signal analysis in python," in Proceedings of the 14th python in science conference, pp. 18–25, 2015.

[10]. F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, et al., "Scikit-learn: Machine learning in python," No. Oct, pp. 2825–2830, 2011.

[11]. "Sound Properties Amplitude Period Frequency Wavelength." http://mx.up.edu.ph/cgi-bin/read.php?article=sound+ properties+amplitude+ period+frequency+wavelength&code=989ae161d9a35c1183f1801a5f915c97.

[12]. "Emotion"https://artsandculture.google.com/entity/emotion/m02tjx?hl=en

[13]. "What the Hell is Perceptron? - Towards Data Science." 09 Sept. 2017, https://towardsdatascience.com/what-the-hell-is-perceptron-626217814f53.

[14]. Gaurav Sahu,"Multimodal Speech Emotion Recognition and Ambiguity Resolution",arXiv:1904.06022v1 [cs.LG] 12 Apr 2019

[15]. Radhey Shyam. Convolutional Neural Network and its Architectures. Journal of Computer Technology & Applications. 2021; 12(2): 6–14p.