Published by :
http://www.ijert.org

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 10 Issue 11, November-2021

Analysis of Vocal Pattern to Determine Emotions using Machine Learning

Mrs. Veena Potdar[1], Mrs. Lavanya Santosh[2], Supritha M Bhatt[3], Yashaswini M K[4],
[1]Associate Professor, Department of Computer Science and Engineering,
Dr. Ambedkar Institute of Technology, Bengaluru, India
[2]Assistant Professor, Department of Computer Science and Engineering,
Dr. Ambedkar Institute of Technology, Bengaluru, India
[3,4]Students, Department of Computer Science and Engineering,
Dr. Ambedkar Institute of Technology, Bengaluru, India

***Abstract*:- The potentiality to vary vocal sounds in order to produce speech is one of the major features which sets humans apart from other living beings. We can categorize human emotion by several attributes such as pitch, timbre, loudness, and vocal tone. It has often been noticed that humans express their emotions by varying different vocal attributes during speech generation. Hence, identification of human emotions using voice and speech analysis has a practical possibility and could potentially be beneficial in improving human conversational skills. One can follow an algorithmic approach for detection and analysis of human emotions with the help of voice and speech processing. The proposed approach has been developed with the objective of incorporation with futuristic machine learning systems for improving human-computer interactions with the help of machine learning models SVM (Support Vector Machine) and CNN (Convolution Neural Network) by the extraction of MFCC features. The above-mentioned models are trained using RAVDESS and TESS datasets.**

Keywords: ***SVM, CNN, MFCC, RAVDESS and TESS datasets.***

I. INTRODUCTION

Emotions are very important for humans in order to impacting perception in their everyday activities such as communication, learning and decision- making. They are communicated through speech, facial expressions, gestures and other non-verbal actions. Human emotion recognition plays a major role in the interpersonal relationship. In recent times automatic recognition of emotions has been an active research topic. This encourages several advancement activities in this field. Emotions are produced in the form of speech, hand and gestures of the body and facial expressions. Hence extracting and understanding emotions has a high importance in order to enhance interaction between human and machines.

Emotion recognition is the process of recognizing human emotions. The human voice is very multifaceted and carries a multitude of emotions. Emotion in speech can be used to get additional insight about human behavior. If we further analyze, we can better understand the intentions of people, whether they are unhappy clients or cheering fans.

II. EXISTING SYSTEM

Previously Speech Emotion Recognition was implemented using other methodologies and datasets. Some of them make use of different kinds of neural networks and different types of classifiers for emotion classification. Some of the paper [4] have used different datasets like Berlin and Spanish Speech databases.

Berlin Emotional Speech Database: The Berlin database is extensively used in emotional speech recognition. It has 535 statements spoken by 10 actors in which 5 are female and 5 are male mimicking 7 emotions namely anger, boredom, fear, joy, sadness, disgust, and neutral.

Spanish Emotional Database: The INTER1SP Spanish emotional database has statements from two professional actors (one female and one male speaker). The dataset was recorded two times in the 6 basic emotions, even neutral emotion is also recorded (anger, joy, fear, sadness, disgust, surprise, neutral). In addition to it 4 neutral variations such as soft, loud, slow, and fast recordings can be found in the dataset.

Jerry Joy paper [2] proposed the use of Multi- Layer Perceptron Classifier.

Multi-layer Perceptron Classifier: (MLP Classifier) depends on fundamental Neural Network to perform classification. MLP Classifier uses a Multi-Layer Perceptron (MLP) algorithm and trains the Neural Network following Back propagation.

To build the MLP Classifier they followed steps mentioned below.

- Defining and instantiating the required parameter to initialize the MLP Classifier.
- Training Neural network by feeding data.
- The trained model is used to make predictions on new (or test) data.
- Calculating the accuracy of the predictions

In the paper published by [1], they have used Random Forest and Decision Tree. But we have trained both SVM and CNN model using the RAVDESS and TESS datasets

Leila Kerkeni paper [4] uses the Spanish database, having the feature combination of MFCC and MS using RNN with the recognition rate 90.05%.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 10 Issue 11, November-2021

Recurrent Neural Networks: (RNN) are best used for learning time series data. While RNN models are good at learning temporal correlations, they suffer from the vanishing gradient problem which increases as the length of the training sequences increases. To resolve this problem, LSTM (Long Short- Term Memory) RNNs were proposed by Hochreiter et al (Sepp and Jurgen, 1997) which uses memory cells to store information so that it can exploit long range dependencies in the data (Chen and Jin, 2015).

Unlike traditional neural network that uses different parameters at each layer, the RNN shares the same parameters across all steps.

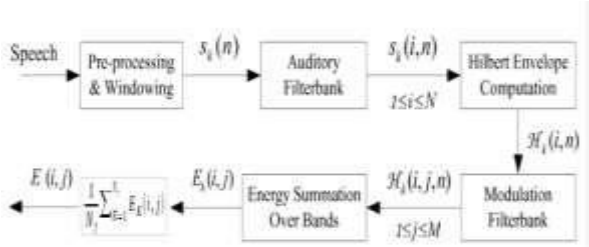

Figure 2: Process for computing the ST representation (Wua et al., 2011).

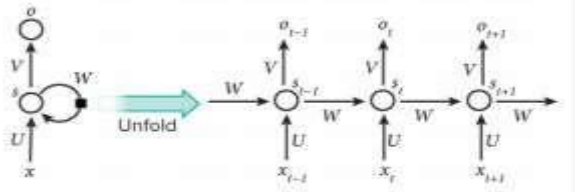

Figure 1: A basic concept of RNN and unfolding in time of the computation involved in its forward computation (Lim et al., 2017).

The hidden state formulas and variables are as follows: st = f(Uxt +Wst−1) (1) with:

- xt , st and ot are respectively the input, the hidden state and the output at time step t;
- U, V, W are parameters matrices.

Dataset	Feature	Average (avg)	Standard deviation (σ)
Berlin	MS	66.32	5.93
	MFCC	69.55	3.91
	MFCC+MS	58.51	3.14
Spanish	MS	82.30	2.88
	MFCC	86.56	2.80
	MFCC+MS	**90.05**	1.64

Table 1: Recognition results using RNN classifier based on Berlin and Spanish databases

III. PROPOSED SYSTEM

In comparison to the existing system, we are trying a different approach by using different machine learning models such as SVM and also neural network models called CNN. Our predictions are way different from the predictions of the existing system. Emotion prediction finds an application in emotional hearing aids for people with autism; detection of an angry caller at an automated call center to transfer to a human; or presentation style adjustment of a computerized e-learning tutor if the student is bored.

In this proposed system, we have used the RAVDESS and TESS datasets.

They are recorded in English which is most widely spoken all over the world. Our dataset has more files when compared to Berlin dataset so that we can train our model using large number of audio files.

RAVDESS: Dataset RAVDESS has 1440 files i.e, 60 trials per actor recorded by 24 actors = 1440 (I.e., 60*24). The RAVDESS is recorded by 24 professional actors comprising of 12 female, and 12 male, vocalizing two lexically matched statements in a neutral North American accent. Recordings expresses emotions like calm, happy, sad, angry, fearful, surprise, and disgust expressions. Each expression is constructed in two different levels of emotional intensity namely normal, and strong, with an additional neutral expression.

TESS Dataset: TESS Dataset has stimuli modeled on the Northwestern University Auditory tests. A collection of 200 different words were audio taped using the carrier phrase "Say the word_' by two actresses aged 26 and 64 years, recordings were made of the set expressing each of 7 emotions namely happiness, pleasant surprise, neutral, anger, fear, disgust, and sadness. We have used 1400 recorded files of older actor in our ML models. So, in total we have used 2840 files to train our Machine learning models.

Working of the system:

Feature Extraction: Dataset in the google drive is taken by librosa.load() method to give 40 MFCC features back for every audio file. These features are added to X(independent) and we save the file name into the y(dependent) variable. Later in SVM and CNN models they are split into X_train and y_train.

SVM implementation: The extracted feature is scaled using standard scalar library and data is split into training and test data respectively. Then we build the svm model using svc library of sklearn.svm as follows:

Model=svc(kernel='linear',random_state=0)

Then we train the model using the fit() method passing X_train and y_train.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 10 Issue 11, November-2021

CNN implementation: The extracted features are converted to numpy array format and we create CNN model with 1 Convolution layer,1 activation='relu' layer, 1 dropout layer, 1 flatten layer and then the dense layer, softmax activation layer. Then we fit X_train and y_train and use predict_classes() to make predictions.

Now these models are put into pickle file or .h5 file so that they can be used for quick predictions.

We have extracted the features from the RAVDESS and TESS datasets especially the MFCC features. We have trained both SVM and CNN model using these datasets. Both are having good accuracy. We have tried to predict the emotion by uploading the audio file to the designed Web page. The model and front-end is been connected using the flask framework.

Emotion prediction:

Once you upload the proper audio file of .wav format and click on the Submit button the emotion predicted will be displayed below the Submit button. This emotion may vary from (Happy, Neutral, Sad, Calm, Angry, Fearful, Surprise, & Disgust).

Accuracy:

The automatic emotion recognition system proposed with the help of SVM and CNN models have the accuracy of 65% and 81% respectively.

	precision	recall	f1-score	support
1	0.89	0.66	0.76	73
2	0.80	0.89	0.84	124
3	0.72	0.83	0.77	129
4	0.72	0.80	0.76	106
5	0.81	0.94	0.87	108
6	0.88	0.76	0.81	139
7	0.86	0.76	0.81	134
8	0.90	0.83	0.87	139
accuracy			0.81	952
macro avg	0.82	0.81	0.81	952
weighted avg	0.82	0.81	0.81	952

Figure 2: SVM accuracy

```
from sklearn.metrics import classification_report
report = classification_report(YTestAry,y_pred)
print(report)
```

	precision	recall	f1-score	support
01	0.78	0.77	0.77	56
02	0.53	0.66	0.59	35
03	0.51	0.57	0.54	83
04	0.64	0.65	0.65	80
05	0.80	0.76	0.78	93
06	0.53	0.68	0.60	62
07	0.70	0.71	0.70	65
08	0.74	0.50	0.60	92
accuracy			0.65	568
macro avg	0.66	0.66	0.65	568
weighted avg	0.67	0.65	0.65	568

Figure 3: CNN accuracy

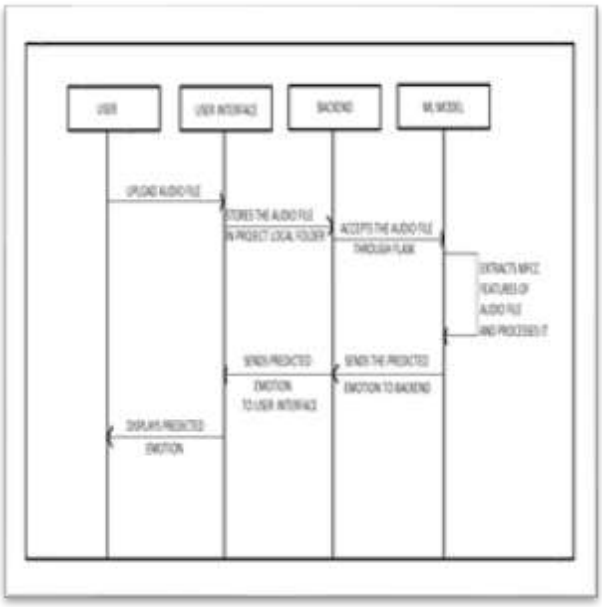

Figure 4: Working of Speech Emotion recognition system

IV. RESULTS

The input files used are the audio files of the trained datasets.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 10 Issue 11, November-2021

The following are the result of the conducted experiment.

Test case	Test case description	Input (Audio file)	Expected output	Actual output	Results
1	Checking for prediction on neutral emotion	03-01-01-01-01-01-13.wav	Displays Neutral	Neutral	Pass
2	Checking for prediction on calm emotion	03-01-02-02-02-01-08.wav	Displays Calm	Calm	Pass
3	Checking for prediction on happy emotion	03-01-03-02-01-02-08.wav	Displays Happy	Happy	Pass
4	Checking for prediction on angry emotion	03-01-05-02-02-02-24.wav	Displays Angry	Angry	Pass
5	Checking for prediction on disgust emotion	03-01-07-02-01-02-16.wav	Displays Disgust	Disgust	Pass
6	Checking for prediction on sad emotion	03-01-04-02-01-02-18.wav	Displays Sad	Sad	Pass
7	Checking for prediction on fearful emotion	03-01-06-01-02-02-16.wav	Displays Fearful	Fearful	Pass
8	Checking for prediction on fearful emotion	03-01-08-01-01-02-04.wav	Displays Surprise	Surprise	Pass
9	Checking for prediction on any emotion	Upload the audio file of any emotion other than .wav extension	Displays an alert message	“Only .wav file allowed”	Pass

Table 2: Result of the conducted experiment.

V. CONCLUSION

The proposed system successfully predicts the emotions of the audio files most of the times. So, this can be further enhanced and used for the criminal case investigations, call centers and even to detect the emotions of any individual automatically rather than predicting the emotion by the other person. Sometimes predicting the human emotion manually by another person is difficult task due to various reasons like (the absence of that person), so usage of this automatic emotion recognition system plays an important role to predict the emotions by saving time as well as money.

VI. FUTURE SCOPE

In the proposed system we have just created the user interface having the option to upload the audio file, further we can enhance its improvements by providing a microphone which directly records the audio and can predict the output, also we can implement the code which converts any format of audio file to .wav when the file is uploaded to predict the emotion of that audio file. The machine learning model can be trained by more datasets to enhance its accuracy.

VII. REFERENCES

[1] Fatemeh Noroozi, Tomasz Sapinski, Dorota Kaminska, Gholamreza Anbarjafari, Survey on vocal-based emotion recognition, Springer Science + Business Media New York 2017.
[2] Jerry Joy, Aparna Kannan, Shreya Ram,
[3] S. Rama, Survey on Speech Emotion Recognition using Neural Network and MLP Classifier, SRM Institute of Science and Technology, Vadapalani Campus, Chennai, India.
[4] S. Lalitha, Abhishek Madhavan, Bharath Bhushan, Srinivas Saketh, Survey on Speech Emotion recognition, Published in: 2014 International Conference on Advances in Electronics Computers andCommunications.
[5] Leila Kerkeni, Youssef Serrestou, Mohamed Mbarki, Kosai Raoof and Mohamed Ali Mahjoub, Survey on Speech Emotion Recognition: Methods and Cases Study, University of Maine, Le Mans University, France, LATIS Laboratory of Advanced Technologies and Intelligent Systems, University of Sousse, Tunisia, Higher Institute of Applied Sciences and Technology of Sousse, University of Sousse, Tunisia.