Predicting The Emotions Based on Emoji's and Speech Using Machine Learning Techniques

Dr K. Lavanya[1], Ch.Lalitha Devi[2], M.Divya Sree[3], P.Nagul Shareef[4]
[1]*Associate professor Lakireddy Bali Reddy College of Engineering, India*
[2,3]*Information Technology, Lakireddy Bali Reddy College of Engineering, Vijayawada, India*
[4]*Information Technology, Lakireddy Bali Reddy College of Engineering , Guntur, India*

Abstract **- Speech consists of assorted information, like language, emotions, what type of message to be communicated with others etc. Emotions are the part of human life in every situation, sometimes one get angry, sad, happy based on the dialogues and behavior of the opposite person. In this work, we have a tendency to square measure aiming to predict the emotions supported the audio files. At first the dataset encompass audio files. Here the emotions typically represented as happy, sad, surprised, angry etc., and could be divided into 2 varieties like positive emotions and negative emotions. Here emoji's are used to predict the emotion of the person, so that it can be quickly identified, for every feeling there'll be a revered emoji format supported that we have a tendency to square measure able to get emoji's for the required emotions given within the datasets. Before applying ways or models on the dataset, feature extraction plays a big role during this speech feeling prediction. Afterward we have a tendency to square measure applying Machine Learning Techniques such as Decision Tree, MLP classifier, neural networks and Augmenting the information using noise injection with Laplace and logistic distribution and pitch shifting and trimming the data so as to induce sensible performance.**

Index Terms **- Feature-extraction, Emoji's, Decision Tree, MLP-Classifier, CNN model, Augmentation methods.**

1.INTRODUCTION

Emotion prediction from somebody's speech is a beautiful field of speech waveform signal methods. It's drawing loads of attention among the applications where recognition of feeling eases the identification and mental standards, like frustration, defeat, surprise, health care, and medicine, etc. As these emotions play an important role while communicating with another person, the detection and analysis of a similar are the significant importance in today's digital world of remote communication. Feeling prediction might be a tough task, as a result of emotions area units are different with different person expressions. There's no common accord on a way to live or categorize them. Extracting the feeling of the person's voice is termed as speech feeling recognition. The speech feeling prediction entails reading the voice indicators to one-of-a-kind the proper feeling with emoji primarily based totally on functions like pitch, frequency, amplitude, structure.

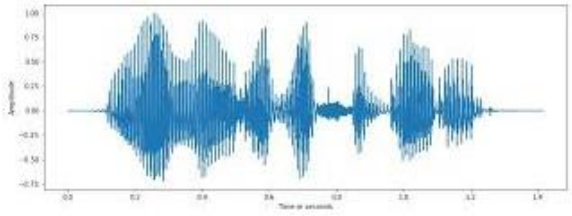

Fig1- Audio signal file wave format

Considering two different datasets such as the RAVDESS and SAVEE datasets to predict emotions in the form of emoji. Each data consist of several audio files in which we considered only five emotions (Happy, sad, calm, fear, angry) from each dataset and considered the neutral emotion. Different types of classification methods are available like Decision Tree, Multi-layer Perceptron classifier, and Convolution neural networks.

Data Augmentation is one more point in the study of Audio files because there may be some disturbance in the data so that here we used Noise injection with Laplace and logistic distribution and Trim & pitching each audio sample. The study of this work, is about predicting emoji's based on emotions and measuring the overall accuracies of different algorithms.

2. EXISTING SYSTEM

Speech recognition system extracts the data using different feature extractions and algorithms to predict

emotions. At first feature extractions are only used to recognize the emotions later applied different algorithms in order to predict the emotion of the person.

In existing system, the predictions are in the form of name of the emotion that are available in the sample datasets. Mel frequency cepstral coefficient, spectrograms, Chromograms etc. are some of the features extracting techniques are used which will extract features from the audio samples.

After applying algorithms emotions are predicted.

3. PROPOSED SYSTEM

In the proposed system, Ravdess and SAVEE datasets are taken for predictions and added emoji's in the place of emotions and extracted the features using Mel power spectrograms which then converts to log scale decibel and also the zero-crossing rate for each audio file using the LIBROSA library. We labeled the data into positive, negative, and neutral by considering the five different emotions based on gender. The feature extractions are drawn for some emotions in figure 2 and figure 3. Based on the time and frequency in Hz, it will store the mean values of each audio file. Typically the number of features depends upon the time duration of the audio wave file. The algorithm takes the input shape of train data with an activation layer of 'relu' and 'softmax', a dense layer of 4 and 3 concerning the predicting classes, and a dropout layer with 25%. Using augmented method with noise Laplace &logistic distribution to the train data and also the audio effect with time and frequency effects with pitch_shifting and trimming the data. Finally passing this augmented data with our CNN model and predicting the labels based on the Train and test data. At the end comparing the accuracy of the predicted values with applied algorithms these are shown in Table.3 and Table.4.

4. FEATURE EXTRACTION

In this, we used the Mel spectrogram converts to log scale decibels and Zero-crossing rate of an audio file using LIBROSA library. The frequencies are converted to the Mel scale with decibel conversion. Considering nmels with 128 where it then converted into power-to-dB using librosa library and stored all the features in a horizontal format considering the mean values of the spectrograms. Zero Crossing Rate describes the audio is present or not and calculates in form of negative, positive,zero. we take features of each file at an offset value of 0.4 by considering the seconds with 3 and 2.5 sec.

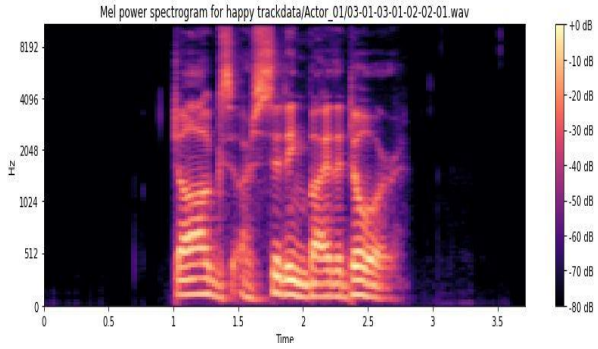

Fig.2.positive track

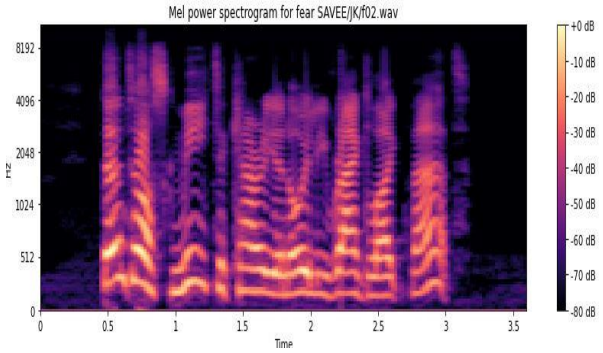

Fig.3.positive track

5. METHODOLOGY

In this paper, Extracted the data by considering the audio files format and labeled according to the required prediction format.

Here, we considered two audio files datasets named Ravdess and SAVEE. Ravdess dataset consists of total of 7356 files of both male and female data in which 2452 files are used for predictions according to the model and total of 1560 files are considered by combining different emotions into positive and negative by gender. The SAVEE dataset consists of initial 480 files with only male dataset, in which 420 files are taken for predicting the emotions.

Applied the machine learning techniques on the datasets, as there are many algorithms, we considered the below algorithms. Each emotion needs to predicted with specific emoji label data, as it is not easier to predict emotion with better accuracy.

For this reason, here we've a bent to pick to match the performance of varied classification ways. For all the methods we've got split the information into 75%,25%, and 60%,40%. Refer Fig.5. and Fig.6. for label emotions.

6.ALGORITHMS

Decision Tree.

In this study, we used decision making on the train and test data with 60% and 75% on train data and tested on 40% and 25% of the dataset's test data. Here we got the accuracy of 52% and 48% for [60:40] whereas for [70:25] we got around 42% for both datasets. Consider that the data is in required format.See that the data must not contain null values

Multi-layer perceptron classifier:

MLP classifier is one of the classification methods which is simpler to the neural network shape. We considered the activation layer with 'logistic', with a batch size of 256,adaptive learning rate is taken and predicted with an accuracy around 46% and 43%.

Convolution Neural Network:

CNN is an efficient algorithm that is widely used in pattern recognition, image processing and voice analysis.IT has many features such as simplicity, fewer training parameters, and adaptability. There are many hidden layers in this CNN-based model. At first, we need to normalize the data such that it converts the object type into numerical values which are used at the time of prediction. Change the dimension of the train and test data. CNN consists of many hidden layers such as dropout, activation, normalization, dense layer, optimizers such as Adam, etc. these are the basic building blocks of neural networks. In this, we used the Conv1D with a batch size of 256, and input shape as train data shape at first, then added the activation 'relu' to the model, Dropout layer with 0.25(25%) and batch Normalization, Max pooling with pool size 8 were it downsamples input representation by taking maximum value over a pool size. The above all metrics used can be refer in Fig4.Callbacks: The objects perform actions at different stages of training such as the start or end of an epoch. ReduceLROnPlateau is used with monitor 'val_loss' at a factor rate of 0.9.It reduces the learning rate when the metric has stopped improving, if there is no improvement seen then the learning rate is reduced. Checkpoints are used to save the model when at every time which can be saved to JSON file and can be accessible to load the weights. Finally, validate the model with test data to predict the labels emojis.

Layer(type)	*Output Shape*	*Param #*
Conv1d_80(Conv1D)	(None,432,256)	2304
Activation_90(Activation)	(None,432,256)	0
Conv1d_81(Conv1D)	(None,432,256)	524544
Batch_normalization_20	(None,432,256)	1024
Activation_91(Activation)	(None,432,256)	0
Dropout_20(Dropout)	(None,432,256)	0
Max_pooling_20(Maxpooling)	(None,54,128)	0
Conv1d_82(Conv1D)	(None,54,128)	262272
Activation_92(Activation)	(None,54,128)	0

Fig.4. CNN Model

Data Augmentation:

Augmenting data for audio files means, it gives better result if there is any noise in the data and removes or changes the frequency of the file by applying some of the augmenting methods such as time,pitch,etc.,Train the model with appropriate augmented methods what need to be considered.

As, synthetic data is generated for an audio, by considering the augmented methods such as noise injection, pitch and time shifting and tuning the speed. In python, the package numpy provides us a simple way in order to handle these methods, whereas Librosa helps us to manipulate voice control over pitch and speed.

we applied Noise injection with Laplace and logistic distribution and time frequency pitch shift and trimming the audio data.

Laplace Distribution with noise injection:

It is similar to Gaussian/normal distribution but has fatter tails, it shows the difference between two independents, identically or similar distributed random variables. It takes the size of the samples, here we calculated the noise with an offset value of 0.05. Below is the formula for Laplace distribution.

f(x,location,scale)=exp(-abs(x-location)/scale)/(2*scale)

Logistic Distribution with noise injection:

When data is passed to the Logistic function, it takes the shape of the samples and drawn out the parameterized logistic values.

Both considers Location, scale, size, where location takes the float values default is zero, scale should be greater than zero, size takes the shape of the data. In this used combination of both Laplace and Logistic distribution.

Probability(x) =
exp((location-x)/scale)/(scale*(1+exp((location-x)/scale))**2)

Time frequency Pitch-shift and Trim:

It shift the pitch tune into nsteps where it takes the time series data of ndarray, with sampling rate and nsteps which is double the pitch change and bins per octave with 12 as default.

Since audio data consists of some unknown noise, due to this the voice over the audio may not be clear so that here trim is an audio effects which trims the silence or noise from the audio files. This can be loaded from librosa library.

Therefore, both the above methods provides synthetic data where we need to combine with original data for better result pre-processed the data by considering initial CNN model.These combinations can be seen in Table.1. for dataset1-Ravdess, Table.2. for dataset2-SAVEE.

By observing Table.1 and Table.2, Accuracy column describes the accuracies of data is divided into 75:25 and 60:40 as train and test data. method1, we got highest accuracy with 57% when considering 75% as training and validated on test data about 20% of dataset.54.5% is seen as highest accuracy for method2, while considering with 60% of train data against test data.

Methods	Applied distribution and effects	Size of dataset	Features Extracted	Accuracy
Method-1(Noise)	Laplace and Logistic	2896	518	45.31% 57%
Method-2(pitch Trim)	Pitch shifting with trim	3008	518	54.5%, 52%

Table.1. Augmented data for dataset1-Ravdess.

Methods	Applied distribution and effects	Size of dataset	Features Extracted	Accuracy
Method-1(Noise)	Laplace and Logistic	672	432	55.01% 54%
Method-2(pitch Trim)	Pitch shifting with trim	672	432	67%, 62%

Table.2. Augmented data for dataset2-SAVEE.

In table.2, for method1 we got the highest accuracy with 55.01% when considering 60% train data against test data.67% is seen as the highest accuracy for method2 while considering 60% of train data against test data.

Emoji Label Count: Count of labels taken for each dataset.

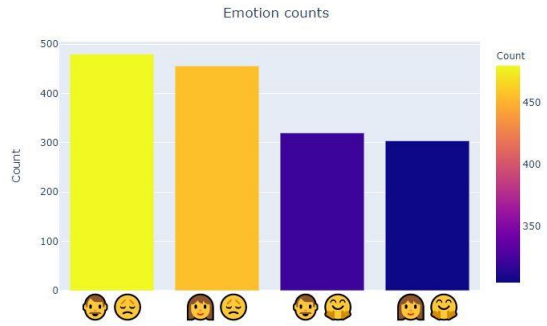

Fig.5. Emotion count with emojis for Ravdess dataset

In Fig.5. Tells about the overall 1560 files we considered for dataset1-Ravdess, here 0 is maleNegative,1-femalenegative,2-malepositive,3-femalepositive.

In Fig.6. Tells about the overall 420 files which we considered for dataset2-SAVEE, here 0-malenegative,1-maleneutral,2-malepositive.

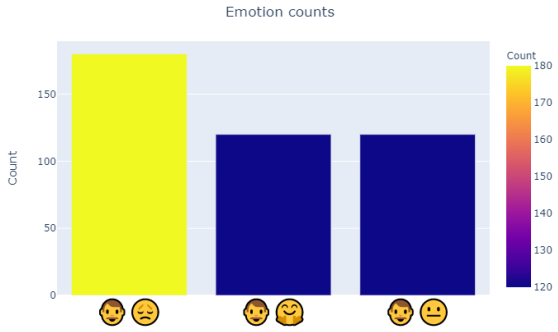

Fig.6.Emotion count with emojis for SAVEE dataset

7. OBTAINED RESULTS

Apply the algorithms on the training data collected to build a recommendation model. To determine the algorithm's accuracy, use this model on each test data instance. We Draw a confusion matrix for CNN with augmented and find the accuracy of these algorithms. Compare the accuracy and find the best model for the dataset to provide better results. Here we used algorithms named Decision Tree, MLPClassifer ,CNN

and used augmented data with noise injection with logistic and Laplace distribution, Time frequency effect using librosa with pitch shifting and trimming the noise of the data.

Here we got better result for our second set of data with an accuracy of 67% with pitchshift and trim augmented data and the other dataset with an accuracy of 57% with Laplace and logistic distribution of noise augmented data.

See below Figure 7, got around 57% as highest accuracy with augmented method1.where as in fig8, represents the accuracy of SAVEE data which was considered, got around 67% accuracy.

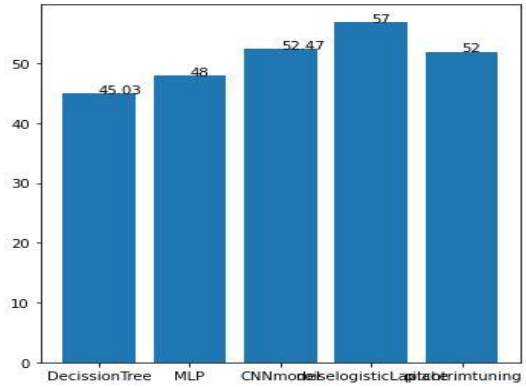

Fig.7. Accuracy of Ravdess data

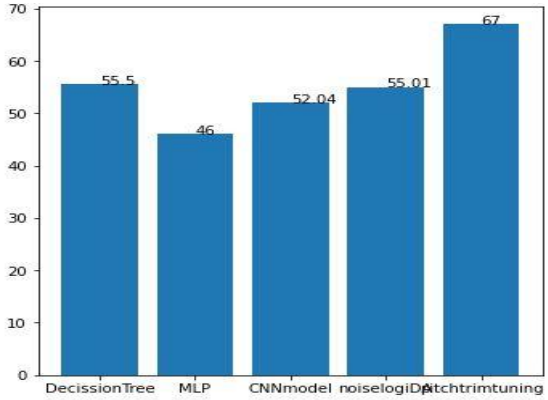

Fig .8. Accuracy of Savee data

8. COMPARISON RESULTS

Comparisons are done for overall data with two different splitting methods for both datasets. Main aim of the work is to understand the better result on different algorithms and augmentation methods which gives a good result at end of this work.

Here, this Table.3. shows the comparison of the results of SAVEE dataset. we got better accuracy for augmented method2 i.e., 67%.

Classification Methods	Accuracy on [60:40]	Accuracy on [75:25]
Decision Tree	52.5%	48.81%
MLPClassifier	45.93%	45.24%
Convolution Neural Network	52.04%	49%
CNN model with Augmented data using 1. Logistic and Laplace noise	55.01%	54%
2.Time frequency shifting and trim effect.	67%	62%

Table.3. comparison of SAVEE dataset

Table.4. shows the comparison of accuracies between both split methods such as 60% and 75% for different methods, we got better accuracy for augmented method1 i.e., 57%, Whereas for Time-Frequency shifting and trim data we got the similar results for both combinations.

Classification Methods	Accuracy on [60:40]	Accuracy on [75:25]
Decision Tree	42.01%	45.03%
MLPClassifier	46.72%	48%
Convolution Neural Network	47.52%	52.47%
CNN model with Augmented data using 1. Logistic and Laplace noise	45.31%	57%
2.Time frequency shifting and trim effect.	54.5%	52%

Table.4. comparison of RAVDESS dataset

9. CONCLUSION

In this work, we applied emoji's to the data in place of emotion which can useful for easiest recognition of an emotion. Here we developed a scalable prediction of emoji's by using algorithms like decision Tree, MLP classifier and neural network with CNN using augmented data. After comparing the accuracy of these, we conclude that augmented method1 with logistic and Laplace, pitch tuning and shifting with trimmed effect on audio data has given better accuracy. This work could be extended by considering Artificial Intelligence and deep neural networks in future.

10.FUTURESCOPE

To increase the overall efficiency of the system, reduce the unwanted background noise in audio files with respect to your model, apply the required factors which are more effective of the augmented methods. By observing the emotion we can predict the exact emoji for the person with the help of Artificial intelligences and deep neural network, we can also

increase the accuracy of the model by using the updated algorithms in the future.

REFERENCES

[1] M. S. Likitha S. R. R. Gupta K. Hasitha and A. U. Raju "Speech based human emotion recognition using MFCC" 2017 International Conference on Wireless Communications Signal Processing and Networking (WiSPNET) pp. 2257-2260 2017.

[2] T. Ozseven "Investigation of the Effect of Spectrogram Images and Different Texture Analysis Methods on Speech Emotion Recognition" ScienceDirect Applied Acoustics pp. 70-77 2018.

[3] K. Tarunika R. B. Pradeeba and P. Aruna "Applying Machine Learning Techniques for Speech Emotion Recognition" 9th International Conference on Computing Communication and Networking Technologies (ICCCNT) 10–12 July. 2018.

[4] Speech Emotion Recognition Using Deep Learning Techniques-19 August-2019.

[5] S. G. Koolagudi and K. S. Rao "Emotion recognition from speech: A review" Int. J. speech Technol. vol. 15 no. 2 pp. 99-117 2012.

[6] H. O. Nasereddin and A. R. Omari "Classification Techniques for Automatic Speech Recognition (ASR) Algorithms used with Real Time Speech Translation" Computing Conference pp. 200-207 2017.

[7] B. Zhang C. Quan and F. Ren "Study on CNN in the recognition of emotion in audio and images" 2016 IEEE/ACIS 15th International Conference on Computer and Information Science (ICIS) pp. 1-5 2016.

[8] M. B. Mustafa A. M. Yusoof Z. M. Don and M. Malekzadeh "Speech emotion recognition research: An analysis of research focus" International Journal of Speech Technology vol. 21 pp. 137-156 2018.

[9] S. Sahu R. Gupta et al. "Adversarial auto-encoders for speech-based emotion recognition" Proc. of Inter-speech 2017