Speech Emotion Recognition Using Convolutional Neural Network (CNN)

[1]Prof. Pranali Deshmukh [2]Mokshada Mahajan [3]Rutika Ovhal, [4]Pooja More

Department of Information Technology
Hope Foundation International Institute of Information Technology, Pune, India

Abstract~Speech is the way that humans can interact with each other also interact with computers, and understanding speech is one of the most important processes that humans communicate. SER's main goal is recognizing the emotion of humans in a given speech. When a human is speaking then analyzing his/her tone and pitch to underlying emotions while using his /her reflected voice. For that, we can use CNN for more efficient results. We use the MFCC for the extraction of features from audio or given extracted input. We use the dataset for evaluating our models and functions. The evaluations of emotions like Happy, sad, angry, neutral, surprised, disgust of speech will defect or we can find it. And using the same we record or predict the ratings of any product bought on the web app which we used to do earlier manually.

Keywords--- Input Extraction, CNN, MFCC, Chroma, Mel Spectrogram Frequency, Librosa, Sound file, NumPy,Scikit-learn,PyAudio, MongoDB, MLP.

Abbreviations--- CNN-Convolutional Neural Network, MFCC- Mel Frequency Cepstral Coefficient, MLP-Multilayer Perceptron, SER-Speech Emotion Recognition.

I. INTRODUCTION

The project includes a web app which has a list of items that user can choose to buy just like a general online store, as we know are made to go through rating process manually but we tried including it in an earlier way, by recognizing emotion through speech which can lead us to know about how the user experience with the product was. And further, the upcoming users can look over the detected ratings and choose wisely. So collecting important information through speech is done by the SER model.

'Multimedia pattern recognition is a rising era that could extract and analyze huge quantities of multimedia information from video and audio assets. In recent years, there was a drastic growth in the software of system studying generation using deep gaining knowledge of to solve diverse popularity troubles. Speech emotion popularity (SER) is an especially large challenge in information the characteristics of speech in media. But, recognizing emotions from the speech is very difficult trouble due to the fact people explicit emotions in unique ways, and the capabilities are uncertain to differentiate the feelings. The paralinguistic hassle is challenging even for humans [1, 5].

CNN is a network that analysis the output to the inputs and finds a pattern between them. This pattern is remembered by the machine. Machine-made to learn the whole scenario and works according to it and process in which machine learns is called training the model, after that the model is trained the machine reacts or behaves to future input and outputs. 'The conventional fashion in speech/audio records retrieval is to recognition on the usage of powerful techniques for semantic evaluation, frequently relying on model selection to optimize the results[2].

'Speech emotion processing and recognition machine is normally composed of three components which might be speech input extraction, feature extraction, and emotion predict. In feature extraction, it usually took the entire emotion audio file as units for function extracting, and extraction contents. Speech is humans speak, and understanding speech is one of the trickiest techniques that the human mind plays. It has been argued that children who aren't capable of apprehending the emotional states of the speakers advanced negative social skills and in some cases, they display psychopathological signs' [3, 4]. The ratings which were given manually earlier are now tried to be predicted.

II. LITERATURE SURVEY

Our prime focus for the literature survey to address the manual rating process to the user experience through speech emotion recognition. Our survey states that the prediction of emotion that reflects or predicts ratings.

The first paper that we studied namely, “Emotion Recognition through Speech Using Neural Network” says the paper aims to enable a very natural interaction among humans and machines. This dissertation proposes an approach to recognize the user's emotional state by analyzing the signal of human speech. Different methods do exist for implementation used on different kinds of datasets and applications.

The second paper we studied was namely, “A comparative study on feature extraction techniques in speech recognition" stating the gaining knowledge regarding all methods to extract useful data used to recognize emotion problem with it was too much calculative task is present if we want to choose the best method which means more time to process and display results.

The next paper studied was “Machine Learning-Based Emotion Recognition using Speech Signal” k Ashok Kumar, J L Mazher Iqbal, stating that reference to more knowledge achieving regarding the procedure like features of voice, uninterrupted voices which can capture interruptionless voice input. But problems found were no help in practical implementation. or the process going to happen is skipped there was the other way of doing it by decoding code or understand or train the interrupted data, or data full of variation in accents and cultures.

The next paper studied was “speech emotion recognition using deep neural network and extreme learning machine” In this paper, they proposed to utilize the deep neural network to extract high-level features from raw data and show that they are effective for SER. They only 20% relative accuracy implement.

The next paper studied was “Speech emotion recognition”The purpose of the emotion recognition system is to use emotion-related knowledge in such a way that human-machine communication will be improved. The important issue with this wasSER system is the signal processing unit in which appropriate features are extracted from the available speech signal and another is a classifier that recognizes emotion from the speech signal.

After the overall study, we tried implementing a system that takes correct inputs, predicts the correct output of emotion in the user, and then predicts or calculates ratings. We tried to increase accuracy compared to previous projects using all the resources like required technologies, methodologies, datasets, databases, libraries, etc.

III. RELATED WORK

Most of the published papers used spectral and prosodic features extracted from raw audio signals. The process of emotion recognition having the extracted features from a related emotional speech selected and then classification is done by these extracted features.

The performance of the classification of the characters such as a combination of MFCC acoustic features. Zamil et al also a used spectral characteristic which is the 13 MFCC obtained from audio input in this SER to classify the 7 emotions using logistic model tree algorithm and getting 70% accuracy [9].

In these, all traditional papers have been published deep neural networks into their project with accurate results. Many authors accept that the most important audio characteristics to recognize emotions are spectral features, MFCC, SVM, FBI.

In another paper Jianfeng Zhao et al he was used merged CNN and Decision Tree feature for recognizing the emotion from speech, the emotions are recognized with 72% and 63% accuracy [9]. H.M Fayek in [10] used various databases eNTERFACE [11] and SAVEE [12] have 6 and 7 classes respectively with exploring different DNN architecture and taking accuracy rate 60%.

In this paper [14], a multi-task learning model for emotion recognition using the RAVDESS [13] speech dataset, they achieved a 57.14% accuracy recognition rate.

IV. WEB APPLICATION AND DATABASE

- Webpage App :

In our SER project, we used to online grocery web app for taking the input in an audio voice format for display the rating in emotions for developing this web app we used different languages to implementing it. Here we used JavaScript, HTML5, CSS for user Interface and connectivity with database. This language is used for implementing the user interface, so users can friendly with our web app. Users can interact with grocery shopping via our web app and give the rating in the form of voice and store this audio voice in the SER model to predicting the emotions and the web app display rating with predicting emotions.

- Database:

MongoDB

MongoDB is a report-type database which means is storing the data in JSON-like documents. We trust this is the natural way to think about data. MongoDB is more expressive and strong than the traditional row/column model, each database contains a collection which is the documents. The size and contains particular documents can be different from each other.

MongoDB having a data model allowing to representing the hierarchical relationships to storing the arrays, and other more difficult structures more simply. MongoDB is very scalable we can easily run hundred plus nodes with millions of documents.
Here, In our project, we used this strong database to store the users' information and rating given by users in our database.

The rating we used in our project which is audio format are happy, peaceful, wonderful, excellent, awesome, just wow, nice, very good, satisfaction, delightful, good, ok, average, decent, moderate, not good, worst, bad, poor, useless, horrible, hated it, waste of money, rubbish, etc.

I. PROPOSED METHODOLOGY FOR SER

A. Input Extraction

Input Extraction means our input in the form of audio which is evaluating using the MFCC feature.

Dataset:

There are two main datasets are used to SER RAVDESS and SAVEE, we used RAVDESS to evaluating our models. In RAVDESS there are two types of data speech and song.

We used the RAVDESS dataset for our system. The RAVDESS is Ryerson Audio-Visual Database of Emotional speech and song. It contains 7356 files the size is 24.8 GB. This database contains 24 professional voices (12 female, 12 male). It is provided by the Kaggle. It can use emotions like Happy, sad, angry, disgust, surprise these are various speech emotions.

We take a sample audio file for the work using IPython and display it with python's librosa library then we take an audio file which is a Male actor speaking in a neutral tone then we plot the wave of this file using librosa.display.waveplot:

Wave plots plot the amplitude of the signals over time using the MFCC, log Mel spectrograms feature extraction model to use for modeling. Feature extraction is important for modeling, therefore, it can convert an audio file into a format that can understand our models.

B. CNN And MLP

- CNN Algorithm

Convolutional Neural Network (CNN) is used to separating the emotions and predicting result with its accuracy. CNN has their "neurons" arranged greater like the ones of the frontal lobe, the area responsible for processing visual stimuli in people and different animals. CNN uses a system such as multiplayer perceptron that has been created for reducing the process requirement CNN has main three layers. An input layer, an output layer, and a hidden layer. They include other sub-layers multiple convolutional layers, pooling layers, fully connected layers, normalization layers these layers reduce the drawbacks of speech and increase the efficiency of a result of the system that has the most effective easy to train speech or audio voice input in raw pixel data and trains this model then extract with features like MFCC and MFCC automatically apply the classifier MLP.

In our CNN model we have four mostly used primary layers:

1. Convolutional layer: identifies salient areas at durations, length utterances which can be variable and depicts the function map collection.
2. Activation layer: a non-linear activation layer function is used as customary to the convolutional layer outputs. On this, we've got used the corrected linear unit (ReLU) in the course of our paintings.
3. Max pooling layer: this layer enables alternatives with the most costly to the dense layers. It allows maintaining the variable-length inputs to a set-sized characteristic array. This layer is periodically inserted in CNN and its fundamental characteristic is to reduce the size of quantity which makes computation rapid and reduces reminiscence. The two kinds are max-pooling and average pooling.

4. Dense layer

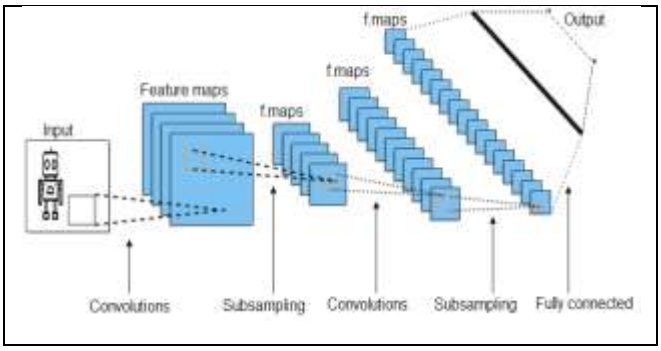
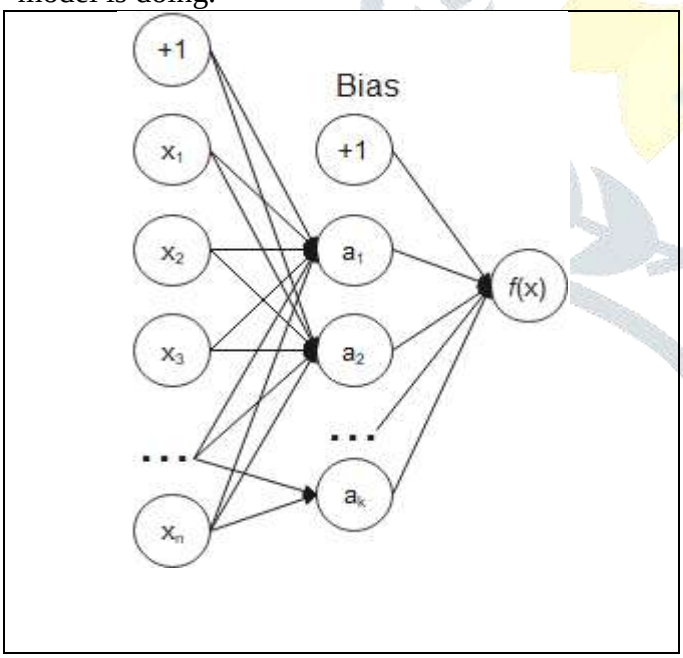


Fig 1.CNN architectures with convolutional, pooling (subsampling), and fully connected layers for the softmax activation function

Algorithm:

Here, we use Convolutional Neural Network (CNN) algorithm for our system speech emotion recognition. Which is used in many modules for recognizing the emotions and classifiers are used for classifying the emotions such as happy, sad, angry, surprise, disgust, neutral. We used Python for implementing the CNN algorithm, here python used various libraries to processing our model. These Libraries are Sound file, Librosa, Numpy, PyAudio, Scikit-learn to extracting audio features and training the model also testing how good our model is doing.



parameter 2.

The input layer has a set of neurons {xi|x_1,x_2,…..x_m} presenting the input feature, each neuron in the hidden layer passes the values from the previous layer with weighted linear summation w_1 x_1 + w_2 x_2+……+w_m.. x_m. The output layer taking the values from the last hidden layer and transforming them into output values. The MLP classifier is used to implement the multilayer perceptron technique,

The algorithm is as follows:

Step 1: First we take the audio speech file through the web application in the form of rating in audio format.

Step 2: This file is plotted waveforms and spectrograms.

Step 3: Then we use the Sound file Library for the read given audio file and then also use the LIBROSA a Python library, we extract the MFCC Mel frequency cepstral Coefficient along with the 10-20.

Step 4: This data processing, we divide this data in train and test after using CNN algorithm and then performing the operations.

Step 5: We check the emotions from the trained data of the human voice.

Step 6: After training the audio voice file, we test the voice file for recognizing the emotions.

Step 7: When the emotion is recognized our system predicts the emotion.

Step 8: Display the predicted emotions with rating.

- MLP Classifier:

enter the unknown test dataset as an input, it will return the parameters and predict the emotions as per training dataset values. The accuracy of the system is displaying in the form of a percentage. Which is the final output of our project.

emotions for tasting execution of the model, if we MLP trains used a given set of training example:

Where

$X_i \in R^n$ and $y_i \in \{0, 1\}$

Where layer,

$$f(x) = W_2 g(W_1^T x + b_1) + b_2$$

One hidden neuron MLP has the function.

Where $W_1 \in R^m$ and W_2, b_1, $b2 \in R$ are model

W_1, W_2 represents the weight of the input layer and hidden layer respectively.

b_1,b_2 represents bias added on the layer and output layer respectively.

which is a feed-forward artificial neural network class (ANN). Back propagation is a training technique used by MLP.

In speech emotion recognition, the input voice is given as the input. The datasets suffering from the number of blocks of processes which is runnable to helping for analysis of speech attributes. The data is pre-processed to updating it to the correct formal and selecting features from the input voice file. Which is extracting using different stages such as framing, hamming, windowing, etc. this process helps in break the file into the number format which presents the frequency-time, amplitude, or other parameters. Which is displays the result for analyzing the audio files.

After extracting the feature from the audio file the model is trained. We use the RAVDESS dataset for an audio file. Which have 24 actors with variations. In parameters for training, we having the numerical values of emotions and their respective features corresponding in various arrays.

This array is given as an input to MLP classifiers that have initialized. The classifier identifies different categories in datasets and classifies them into various

$(x_1,y_1),(x_2,y_2),.....,(x_n,y_n)$

g(.): R → R is an activation function set by default hyperbolic tan.

Given as

$$g(z) = \frac{e^z - e^{-z}}{e^z + e^{-z}}$$

For binary classification, f(x) passes through the logistic function g(x) = 1/(1+e^z) to obtain output values between zero If there is more than two classes f(x) itself having vector size (n-classes). It passes logistic function through the softmax function written as:

$$\text{Softmax}(z)i = \frac{\exp(zi)}{\sum_{1=1}^{k} \exp(zl)}$$

Zi = ith element of input to softmax.

k = no. of classes.

x having a result of the vector containing probabilities.

The cl

$$Loss(\hat{y}, y, W) = \frac{1}{2}||\hat{y} - y||_2^2 + \frac{\alpha}{2}||W||_2^2$$

ass with the highest probability is the output.

MLP uses different loss functions depending upon problem type. This function is used for cross-entropy classification.

$$Loss(\hat{y}, y, W) = -y \ln \hat{y} - (1-y) \ln (1-\hat{y}) + \alpha||W||_2^2$$

For regression, MLP uses a square error Loss function

In gradient descent, the gradient $\nabla Loss_W$ of the loss concerning the weights is computed from W.

i.e.

$$W^{i+1} = W^i - \epsilon \nabla Loss_W^i$$

i = is iteration step

€= Is learning rate with value larger than 0

The algorithm ends when it has a maximum number of iterations, in other term the improvement in loss is below a certain small number.

C. Feature Extraction And Feature Selection

- Feature Extraction :

There are five features MFCC, Contrast, Mels Spectrogram Frequency, Chroma, and Tonnetz. In this, we use all this feature extraction as a speech rather than waveform which contains unnecessary data.

There are 4 capabilities among the ones there is a particular number of coefficients in MFCC used in our model for detection of emotion. And the 5 features are

- MFCC
- Chroma(Pitch)
- MEL Spectrogram Frequency (mel)
- Contrast
- Tonnetz

Mel Frequency Cepstral Coefficient (MFCC)

The recorded speech has to be preprocessed so that useless such things as noise and additional sounds may be eliminated. The next step is to extract the functions from the preprocessed speech. And the maximum vital feature in speech emotion detection is MFCC. Figure 3 represents the pictorial representation of MFCC of a recorded emotion.

The sounds created by a person all through speech receives filtered via the vocal tract. The sound that comes out of the mouth has a particular structure. The vocal tract structure represents a short-term power spectrum. The short-term power spectrum is depicted by MFCC. So MFCC is used in speaker popularity, word to text conversion, and emotion popularity.

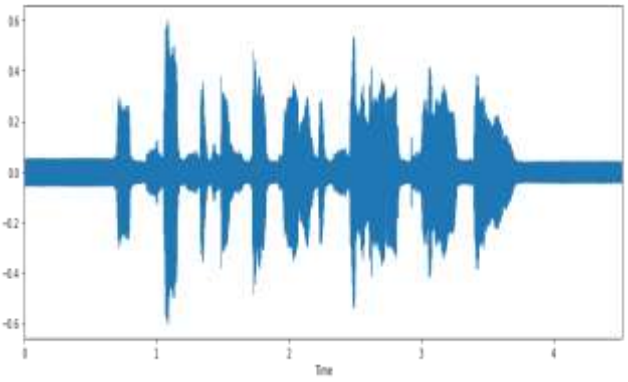


Figure 3. MFCC of an Audio waveform

The mel-frequency cepstral coefficients (MFCC) function extraction approach is the main approach for speech function extraction.

mel (f)=2595 x log 10 (1+f/700)

The diverse steps concerned in MFCC characteristic extraction are:

Chroma(Pitch)

The frequency of sound waves is exactly related to pitch. Frequency is not anything but the range of crest and trough within the sound waves. These crust and troughs represent the vibration of the sound waves. Lower pitch sounds have a minimum variety of vibrations and higher pitch sounds have a big quantity of vibrations. High pitch sounds are very unpleasant to hear and low pitch sounds are the best sounds. Commonly indignant human beings use excessive pitch sounds. Determine 2 represent the blended waveform of different audio files of different feelings. This suggests that kind of that there can be a few differences in pitch for unique feelings.

$Cf(b)=\sum Z-1z=0|Xlf(b+z\beta)||$

MEL Spectrogram Frequency (mel)

The Mel scale connects an unadulterated tone's apparent repeat, or pitch, to its actual recurrence [18]. At low frequencies, people are far better at discerning little variations in pitch than they are at high frequencies. Solidifying this scale makes our features arrange even more eagerly what individuals listen.[19] Contrast.

M(f)=1125 Ln(1+f/700)

Tonnetz

The Tonnetz is a pitch space defined by a network of melodic contribute just inflection links. On a large Euclidean plane, close symphony relationships are shown as short separations.

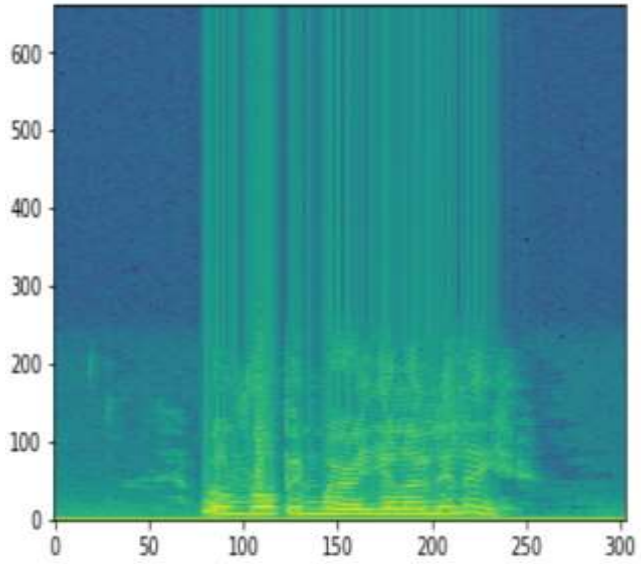


Fig 4.Spectrogram

- Feature Selection:

Filter based feature selection:

- Feature selection is the process of selecting those features that contribute the most to the prediction variable or output that you are interested in, either automatically or manually.
- The filter-based feature selection approach scores the correlation or dependence between input variables, which may then be filtered to choose the most relevant feature.
- We specify some metrics and based on that filter features. Correlation/chi-square is an example of this type of measure.

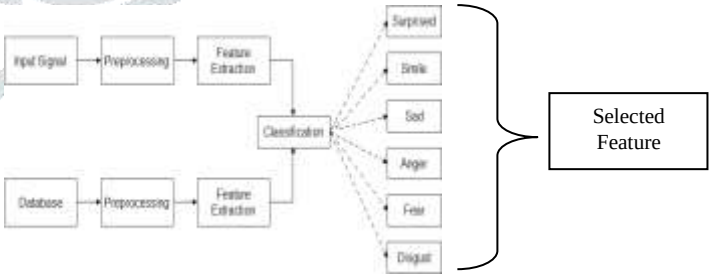
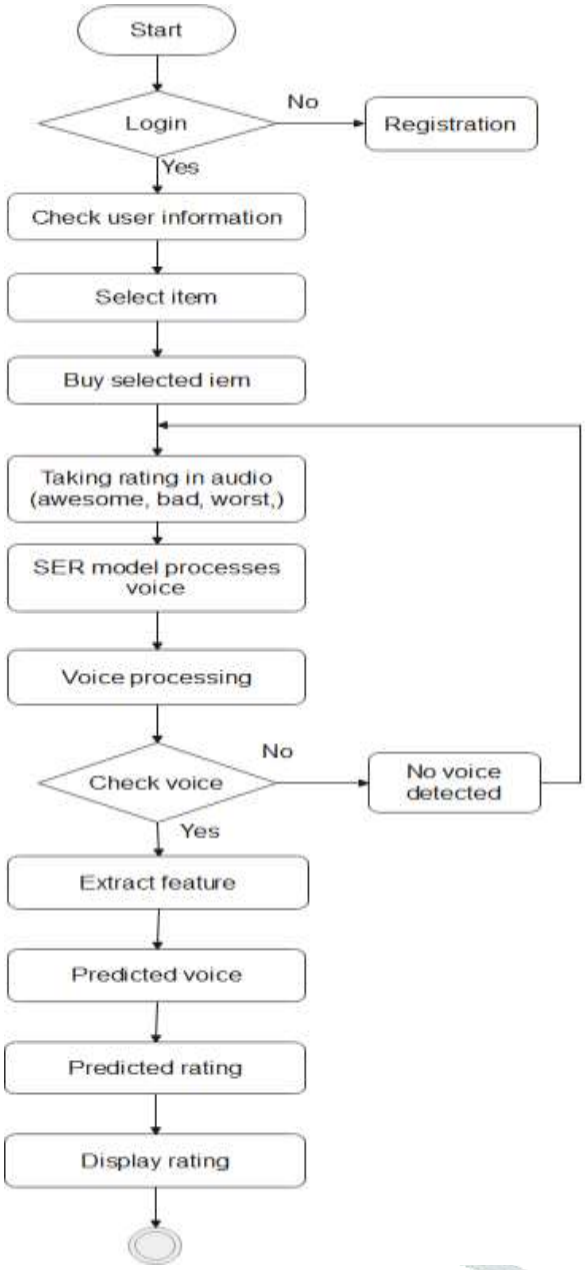


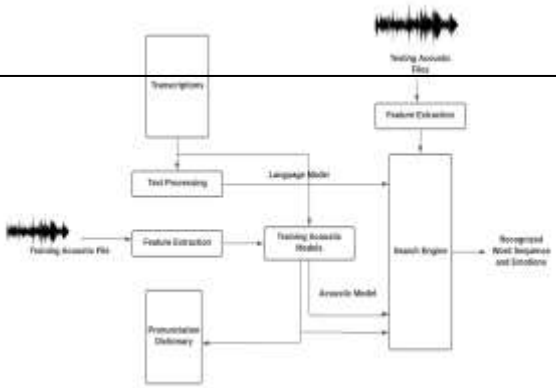
Fig 5. Feature selection

II. IMPLEMENTATION

Flowchart:



The architecture of SER:



Classification report:

Parameter / Emotion	precision	recall	Fi-score	Support
angry	0.85	0.82	0.83	61
happy	0.60	0.66	0.63	41
neutral	0.59	0.62	0.60	21
sad	0.81	0.76	0.78	45

	precision	recall	Fi-score	Support
accuracy			0.74	168
Macro avg	0.71	0.71	0.71	168
Weighted avg	0.74	0.74	0.74	168

Applications:

Speech Emotion Recognition used in the medical field, Customer service in customer service call center conversation used to analyze the behavior of customer calls which is help to improving the customer care service. A recommender system will be useful to recommend the product to the customer. Educational Purpose, Entertainment, for security, lie detection.

III. RESULT

The results for various emotions are captured and tested to achieve more accuracy around 75%. The results are dependent on what emotions the user gives to the machine by which it calculates how much rating is achieved by the web app. The results vary as per the emotion such that if the user's emotion predicted is happy it will calculate the ratings as a 5/5. If the user is unhappy with the experience, the resultant ratings will be less depending upon the calculations. Various emotions are tried upon the SER model to achieve proper outcomes such as correct prediction of emotion and then calculating and display the ratings from it.

IV. CONCLUSION

This paper proposed to rate any user experience through predicted voice which means the system accepts in the form of voice then processed and then predicts by using CNN algorithm, MFCC, Mel, Chroma, Tonnetz, Contrast this features extraction is used to extract emotional characteristics from the emotion speech signal. Our proposed model achieves nearly 75% accuracy using CNN. We use the MLP classifier also for classifying emotion. In the future, we are planning to preprocess our project to remove the time gap between audio files and increasing the accuracy and also the efficiency of the classification of the process.

ACKNOWLEDGMENT

We would like to extend our sincere gratitude towards our project guide, Prof.Pranali Deshmukh for mentoring us throughout project development and implementation by helping us with all the routes to get solutions to problems we faced during implementation with all the valuable knowledge and time.

REFERENCES

1. Zeng, Zhihong, et al. "A survey of affect recognition methods: Audio, visual, and spontaneous expressions." Pattern Analysis and Machine Intelligence, IEEE Transactions on 31.1 (2009): 39-58.
2. Eric J. Humphrey, Juan Pablo Bello, and Yann LeCun wrote "Moving Beyond Feature Design: Deep Architectures and Automatic Feature Learning in Music Informatics." ISMIR, 2012.
3. Monita Chatterjee, Danielle J Zion, Mickael L Deroche, Brooke A Burianek, Charles J Limb, Alison P Goren, Aditya M Kulkarni, and Julie A Christensen. Voice emotion recognition by Cochlear-implanted children and their normally-hearing peers. Hearing research, 322:151-162, 2015.
4. Nancy Eisenberg, Tracy L Spinrad, and Natalie D Eggum. Emotional-related self-regulation and its relation to children's maladjustment. Annual review of clinical psychology, 6:495-525, 2010.
5. Moataz, Mohamed S. Kamel, and Fakhri Karray. El Ayadi, Moataz, Mohamed S. Kamel, and Fakhri Karray. Pattern Identification 44.3, 572-587, "Voice emotion recognition features, classification algorithms, and databases" (2011).
6. Z. Wang and I. Tashev, "Learning utterance-level representations for speech emotion and age/gender recognition using deep neural networks," 2017 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), New Orleans, LA, 2017, pp. 5150-5154.
7. "Speech emotion identification based on HMM and SVM," Yi-Lin Lin and Gang Wei, 2005 International Conference on Machine Learning and Cybernetics, Guangzhou, China, pp. 4898-4901 Vol.
8. Jianfeng Zhao, Xia Mao, Lijiang Chen. Learning Deep features to Recognise Speech Emotion using Merged Deep CNN. IET Signal Process., 2018
9. "Emotion Detection from Speech Signals Using Voting Mechanism on Classified Frames," by Adib Ashfaq A. Zamil and colleagues. 2019 International Robotics, Electrical, and Signal Processing Techniques Conference (ICREST). IEEE, year 2019.
10. H. M. Fayek, M. Lech, and L. Cavedon wrote "Towards real-time speech emotion identification using Deep Neural Networks." IEEE, 2015. "IEEE 9th International Conference on Signal Processing and Communication Systems (ICSPCS)."
11. O. Martin, L. Kotsia, B. Macq, and I. Pitas, April 2006. "The eNtERFACE'05 audio-visible database. In twenty second International Conference on Data Engineering Workshops (ICDEW'06) (pp 8-8). IEEE."
12. Sanjitha. B. R, Nipunika. A Rohita Desai. "Speech Emotion Recognition using MLP", IJESC.
13. "Self Attention Mechanism and Multitask Learning Improved End-to-End Speech Emotion Recognition INTERSPEECH 2019", Graz, Austria, 15–19 September 2019. Li, Y., Zhao, T., and Kawahara, T. INTERSPEECH 2019, Graz, Austria, 15–19 September 2019." In Proceedings of the INTERSPEECH 2019, Graz, Austria, 15–19 September 2019. [CrossRef] [CrossRef] [CrossRef]"
14. Livingstone, S.R.; Russo, F.A. "The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English." PLoS ONE 2018, 13, e0196391. [CrossRef]
15. "Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences", Davis, S., and Mermelstein, P. (1980). Vol. 28 No. 4, pp. "Acoustics, Speech, and Signal Processing, IEEE Transactions on, vol. 357, no. 3, pp. 357-366."
16. "Spoken Language Processing: A Guide to Theory, Algorithms, and System Development", edited by X. Huang, A. Acero, and H. Hon. 2001, Prentice-Hall"
17. Corneliu Octavian Dumitru, IngeGavat, "A Comparative Study of Feature Extraction Methods Applied to Continuous Speech Recognition withinside the Romanian Language," International Symposium ELMAR, 07-09 June 2006, Zadar, Croatia.