Comparison study on speech emotion prediction using machine learning

Reakaa S D and J Haritha

Department of Electronics and Instrumentation Engineering, Bannari Amman Institute of Technology, Sathyamangalam, Erode - 638401

E-mail : reakaa.es19@bitsathy.ac.in, haritha@bitsathy.ac.in

Abstract: Speech Emotion recognition is the process of analyzing human emotions. People vary widely in their precision at recognizing the emotion of others. Support vector machine (SVM) is one of the best algorithms for prediction and handling the voice data. In this speech emotion identification is a difficult task that includes various preprocessing steps. The proposed one, the method is based on the Cepstral Mel frequency Coefficients (MFCC) and voice signal energy as Input and use of the RAVASS emotional database function discourse. In this research work support vector machine algorithm, linear kernel processing is used to predict the human emotions.

In Feature Extraction using Librosa module to handling audio and music type of data. Librosa can extract the values of Mfcc-Mel-frequency Cepstral coefficients it’s used for identifying monosyllabic words in voice recognition. Chroma mainly used for similarity and audio matching. In voice data each data are applied to feature extraction technique. Then generated matrix format of values for calculation. In proposed Support vector machine algorithm for act as a classifier to find emotions. Mfcc, chroma, Mel features are directly send out the Support vector machine (SVM) algorithm. Its take a class of eight emotions like happy, sad, calm, angry, neutral, fearful, disgust. Based on our features (mfcc, chroma, Mel) to classify emotions. The comparison study on SVM and KNN algorithms, SVM has predicted 4% better accuracy than KNN.

Keyword:-speech emotion, feature extraction, MFCC, chroma, support vector machine, emotion.

1. INTRODUCTION

Emotions are plays main role in day by day life. It very useful for rational and intelligent decisions. Emotions are helps to understand the feelings of people and gives feedback to others. Machine learning plays important role on prediction and classifications. In emotion classification SVM was perform main role, SVM is a supervised machine learning algorithm. There are handling numerical data very well. In pre-processing .wav file are convert to matrix format using Librosa features like chroma and mfcc. Sound file module was used to read .wav files. In dataset contains there are 24 folders inside the folders 50 audio files are present. Emotion analysis and sentiment analysis both concept machine learning take a Major contribution. Voice based analysis are previously accessed. Based on frequencies and pitch range we were found the emotions. Previously, voice-based tests are available. Based on the frequency and pitch variety, we’ve been able to find the emotions. Speech Emotion Analysis is complex by the truth that vocal communication is an evolutionarily old non-
verbal mechanism, coded in an iconic and continuous style, which contains emotions and mesh with verbal signals that are coded in an arbitrary and precise style. However, Voice researchers are debating the volume to which asp is verbal and non-verbal.

2. LITERATURE REVIEW

Chunyi Wang’s proposed “Speech Emotion Recognition algorithm based on multi-functional and multi-lingual research to improve accuracy if there are broad datasets and robust acoustic features in speech emotion recognition.

Yi-lin lin, Gang wei(2005) research work focused on the ‘Speech Emotion Recognition on Gender Independent and the Hidden Markov Model (HMM) and the Support Vector Machine (SVM) algorithm to identify five emotional states.

Ashish B. Ingale, D. S. Chaudhari (2012)’ Speech on algorithm identification of emotions and key issues are addressed in this paper to improve accuracy.

Chen Caihua (2019) “The multi-modal emotion detection lacks the explicit mapping the relation between emotion Recognition country and audio, picture features. The extracting is effective emotion detection records from the audio and visible statistics is always a difficult technique. The modelling of blurring and records redundancy is not solved the emotion detection version is often confronted with the trouble of low efficiency.

Nan Song, Hongwu Yang (2019) “For Sign language recognition the help vector machine are used to classify the feelings and facial expression popularity are accrued with the aid of deep neural network. A hidden Markov version based Mandarin – Tibetan bilingual emotion synthesizer is teach by using addresser adaptive training with a Mandarin emotional speech corpus.

M. Albornoz, Diego H. Milone, Hugo L.Rufiner (2017) “Emotional state studying is a major difficulty to achieve greater natural speech interactive systems. The research work is feature is extracted from a time frequency analysis with an acoustic signal.

Margarita Kotti, Yannis Stylianou (2017) “The research work addresses emotion recognition in movie audio track the feature are extracted from the collected acted facial expression in the wild 5.0 dataset. Random forest with support vector machine are implemented for classification. Linear kernel used in the feature extraction to get the efficient accuracy.

Shilpi Gupta, Anu Mehra, Vinay (2015) “Speech Emotion Recognition using SVM with thresholding fusion research presents a methodology to increase the assurance level of emotion classification.

3. PROPOSED SYSTEM

3.1 Algorithms for Speech Emotion prediction

In people’s conversion, voice is the most important form of communication. The emotional state of the addressee is predicted by a speech. Comparison between humans and machines to recognize the state of emotion, it is not difficult for people to understand the voice, but let the software recognize the state of emotion. In order to define the emotional state of the addressee, the computer first pre-processes the addressee’s voice data, extracts the emotional features of the speech signal, and then sets it up. These are the some of the algorithms for speech emotion recognition in machine learning

- Principal component analysis
- K-nearest neighbourhood
- Hidden Markov model
• Support vector machine

In the above algorithms support vector machine and K-nearest neighbourhood algorithm used in this research work. For predicting the accuracy of emotion feature selection support vector machine algorithm is mainly used.

3.2 K-nearest neighbourhood Algorithm
K-nearest algorithm is a continual set of rules that attempts to segregate the dataset into K pre-defined wonderful non-overlapping category in which every data factor belongs to most effective one organization. It tries to make the intra-category information points as associated as feasible even as additionally maintaining the class as distinctive (a long way) as feasible. It attaches records factors to a class such that the sum of the squared distance among the information factors and the class’s centroid is at the minimum. The much less alteration we’ve within series, the greater homogeneous the records factors are in the identical class.

3.3 Support Vector Machine Algorithm
The (SVM) is the supervised device mastering and Classification and Regression Algorithm. It can clear up the liner and non-liner problem the algorithm creates a line or a hyper aircraft. The separates the records into training. Support vector device is one of the maximum famous supervised learning algorithm. The main motive of SVM set of rules. Create the great line or decision boundary that that may divide to the n-dimensional Space into records classes this is we will effortlessly apply the brand-new records inside the right class inside the future. This amazing choice boundary is referred to as a hyper-plane as shown in figure 1.

Hyper-plane is line divides data point into two classes
\[ y = ax + b \]
\[ a = x + b - y = 0 \]

Let vector \( x = (x, y) \) and \( w = (a, -1) \) then in vector from hyper-plane is
\[ W . X + b = 0 \]

3.3.1Types of SVM:
Linear SVM: This is used for linearly separated data that means if data set can be classified can be used to two classes in the single straight line then such data gives to descriptive linearly separated data that is called liner SVM.
\[ f(x) = \text{sign}(w^T \cdot x + b) \]

Non-linear SVM: This is used for linearly separated data that means if data set can’t be classified cannot used by a two classes in the single straight line that kind of data called non-liner SVM.

\[ f(x) = \text{sign}(\sum \alpha y_i K(x_i, x) + b) \]

Polynomial: In the machine learning the polynomial is the kernel function it commonly used to with support vector machine (SVM) in the training samples.

Depending on the variable and then allowing to learn the non-linear data set as indicated in the flowchart shown in figure 2.

Input features are binary-valued like booleans.
- Historical data is classification data.
- SVM data formatting is numerical type.
- SVM parameters is label and features
- SVM training process is used to X_train and Y_train
- Model function.
- V-cross validation is the find out the accuracy value.
- Trained model the used to fit model function in python
- New data is a classification observed value.

![Flowchart of SVM algorithm](image)

**Figure 2** Flow chart of SVM algorithm

### 4. ANALYZING AND PREDICTING THE ACCURACY OF ALGORITHM

**4.1 Process of algorithm**

**4.1.1 Dataset description**
In this research a .wav like audio format file are implemented and each file contains one sentence with different emotions like happy, sad, etc., one folder contains 56 .wav files totally 26 folders are there each .wav files are labeled to emotions. We take eight emotions. Using sound module file read the .wav audio data.

4.1.2 Feature Extraction:
In this proposed model feature extraction plays one major role. In this Librosa module used to extract features from sound files. Features are like matrices values. Features parameters are MFCC, CHROMA, MEL this all are called feature values.

4.1.2.1 MFCC (Mel Frequency Cepstral Coefficients):
Any periodic component (for example, echoes) appears as sharp peaks in the corresponding frequency spectrum in the standard analysis of time signals (i.e. Fourier spectrum. This is obtained by applying a Fourier transform on the time signal). Any Cepstrum function is obtained by applying a spectrogram to Fourier Transform. The unique aspect of MFCC is that it is taken on a Mel scale, a scale that relates to the actual measured frequency of the perceived frequency of a sound. It scales the frequency in order to match what the human ear may hear more closely. The envelope of the speaking signal's temporal power spectrum is representative of the vocal tract and this envelope is correctly represented by MFCC.

4.1.2.2 Spectrogram for MelOn:
Overlapping windowed segments of the signal, a Fast Fourier Transform is computed, and we get what is called the spectrogram. This is just a spectrogram representing the amplitude mapped on a scale of Mel.

4.1.2.3 Chroma:
Usually, a Chroma vector is a 12-element function vector showing how much energy is present in the signal of each pitch class on a typical chromatic scale.

In this process we find this three parameter values using Librosa to each files then all the files are ready to splitting like test and train.

4.1.3 System Design:
The system design flow is shown in figure 3. As shown in figure 3, the audio file is subjected to feature extraction and then split into train and test data. The training data is used to train the SVM algorithm and then the test data is introduced to the trained model for prediction and finally the result i.e., the emotion is detected.

Figure 3. System design

4.1.4 Data Processing:
In this case data was split into two types test and train in the ratio of 20% for analysis purpose train data contains 80% percent of values, test data contains 20% of values train data used to train a SVM model test data was cross check the Algorithm.
4.1.5 SVM algorithm:
After getting train data we implement the SVM algorithm. SVM was taken numerical inputs as features. In this features are metrics format the label are strings like happy sad that all are 70% percent data used to train. Linear kernel was applied here, using vector formula to train the model

\[ f(x) = \text{sign} (w^T x + b) \]

F(x) - training value
X' - features test
W' - features train
B – Label

The metrics values are convert to vectors then applied in the formula that equal emotions also applied to the label field based on that values and labels model get trained after completing training process applied test and evaluation process.

4.1.6 Prediction:
In this field to predict values from test data. Test data also contains same format but compared to train data size was less than 3%. Test data was applied to predict portion here only taken feature not consider labels.

4.1.7 Performance analysis:
Using the test data to predict one output based on that output and original output comparison to understand the performance of model. Here we taken confusion matrix accuracy, f-measure, precision, recall this type of performance score.

5. RESULTS AND DISCUSSION

Speech emotion analysis was major section of Artificial intelligence and Machine learning. SVM algorithm plays important role in this research to classify the emotions based on Librosa features. After analyzing Dataset process then conclude the result SVM gives 60% of accuracy for our process. In future we implement Convolution Neural Network for identify emotions in same data.

The accuracy of SVM algorithm is shown in figure 4, blue wave indicates the actual trained data and the red wave shows the accuracy of predicted test data.

![Figure 4](attachment:image.png)
The accuracy of KNN algorithm is shown in figure 5, blue wave indicates the actual trained data and the green wave shows the accuracy of predicted test data.

![Figure 5](image-url) Accuracy Wave form for KNN algorithm

Figure 6 shows the comparison chart for SVM and KNN Algorithms, in SVM algorithm 59% of accuracy is predicted and KNN algorithm predicted 54%. It shows that SVM gives better accuracy than KNN algorithm.

![Figure 6](image-url) Comparison Chart for SVM and KNN Algorithms

| EMOTIONS | SVM ALGORITHM | KNN ALGORITHM |
|----------|---------------|---------------|
|          | Precision     | Recall        | Precision | Recall |
| Calm     | 55            | 69            | 46        | 81     |
| Disgust  | 58            | 62            | 45        | 53     |
| Fearful  | 72            | 67            | 57        | 59     |
| Happy    | 53            | 50            | 71        | 47     |
| Neutral  | 56            | 53            | 50        | 35     |
| Sad      | 64            | 55            | 68        | 45     |

Table 1 shows the comparison for SVM and KNN algorithms, percentage of precision and recall results of train and tested datasets. The precision percentage is based on how much calm emotion is predicted as calm and recall means it is repeated to predict the calm emotion and it vice versa for all the emotion.
Table 2 Comparison of Accuracies of Precision and Recall

| Stimulus | SVM ALGORITHM | KNN ALGORITHM |
|----------|---------------|---------------|
|          | Precision (%) | Recall (%)    | Precision (%) | Recall (%) |
| Accuracy | 60            | 59            | 56            | 53          |
| macro avg| 60            | 59            | 57            | 54          |

Table 2 shows the comparison of macro and weighted average of SVM and KNN algorithms, precision and recall average is known as macro average and f-1 score and support is known as weighted average.

Table 3. Confusion Matrix of SVM

| Stimulus | Calm | Disgust | Fearful | Happy | Neutral | Sad |
|----------|------|---------|---------|-------|---------|-----|
| Calm     | 18   | 1       | 0       | 0     | 2       | 5   |
| Disgust  | 3    | 21      | 3       | 5     | 1       | 1   |
| Fearful  | 1    | 2       | 18      | 6     | 0       | 0   |
| Happy    | 1    | 9       | 3       | 18    | 4       | 1   |
| Neutral  | 3    | 2       | 0       | 1     | 9       | 2   |
| Sad      | 7    | 1       | 1       | 4     | 0       | 16  |

Table 3 shows that result of SVM algorithm, trained data that predicted accurate emotion like happy as happy 18%, happy as calm 0%, happy as disgust 5%, happy as fearful 6%, happy as neutral 1%, happy as sad 4% and other emotions also presents same as happy.

Table 4. Confusion Matrix of KNN

| Stimulus | Calm | Disgust | Fearful | Happy | Neutral | Sad |
|----------|------|---------|---------|-------|---------|-----|
| Calm     | 21   | 0       | 0       | 0     | 4       | 1   |
| Disgust  | 4    | 18      | 4       | 3     | 2       | 3   |
| Fearful  | 1    | 8       | 16      | 2     | 0       | 0   |
| Happy    | 3    | 8       | 7       | 17    | 0       | 1   |
| Neutral  | 10   | 0       | 0       | 0     | 6       | 1   |
| Sad      | 7    | 6       | 1       | 2     | 0       | 13  |

Table 4 shows result of trained data that predicted accurate emotion like happy as happy 17%, happy as calm predicted 0%, happy as disgust 3%, happy as fearful 2%, happy as neutral 0%, happy as sad 2% and other emotions also presents same as happy.

6. CONCLUSION

In this research presents the methodology for predicting the speech emotion like happy, sad, calm, neutral, fearful etc., based on SVM algorithm. The KNN algorithm predicted 53% of accuracy and
SVM algorithm predicted 59% of accuracy emotions respectively. The comparison between the SVM and KNN algorithms, the SVM algorithm is best for predicting the features of speech emotions. In SVM algorithm, linear kernel processing is implemented and it improves the efficiency of accuracy, precision and recall. In present work SVM algorithm in machine learning is used and for further research work we can analyses the prediction by CNN in deep learning.

REFERENCES

[1] Yi-Lin Lin, Gang Wei, “SPEECH EMOTION RECOGNITION BASED ON HMM AND SVM” Proceedings of the Fourth International Conference on Machine Learning and Cybernetics, Guangzhou, 18-21 August 2005.
[2] Ashish B. Ingale, D. S. Chaudhari “Speech Emotion Recognition” International Journal of Soft Computing and Engineering (IJSCE) ISSN: 2231-2307, Volume-2, Issue-1, March 2012.
[3] Nan Song, Hongwu Yang “A Gesture-to-Emotional Speech Conversion by Combining Gesture Recognition and Facial Expression Recognition” 2018 First Asian Conference on Affective Computing and Intelligent Interaction (ACII Asia).
[4] Chen Caihua “Research on Multi-modal Mandarin Speech Emotion Recognition Based on SVM” 2019 IEEE International Conference on Power, Intelligent Computing and Systems.
[5] E. M. Albornoz, D. H. Milone & H. L. Rufiner “Feature extraction based on bio-inspired model for robust emotion recognition” Soft Computing, Vol. 21, No. 17, pp. 5145-5158, Sep, 2017.
[6] Margarita Kontili and Yannis Stylianoul “Effective Emotion Recognition in Movie Audio Tracks” ICASSP 2017.
[7] Shilpi Gupta, Anu Mehra, Vinay “Speech Emotion Recognition using SVM with thresholding fusion” 2015 2nd International Conference on Signal Processing and Integrated Networks (SPIN).