Speech Emotion Recognition System Using MLP

Vinita Chugh[1],Shivanghee Kaw[2] ,Surabhi Soni[3],Varsha Sablani[4] & Rupali Hande[5]
department of Computer Engineering
Vivekanand Education Society's Institute of Technology
Chembur,India
2018.vinita.chugh@ves.ac.in, 2018.shivanghee.kaw@ves.ac.in, 2018.surabhi.soni@ves.ac.in,
2018.varsha.sablani@ves.ac.in, rupali.hande@ves.ac.in

***Abstract*—Speech is the most natural way of demonstrating ourselves as humans. It is only natural then to expand this communication means to computer applications.In this paper many kinds of emotional speech corpora are similar in expression of speech accession,expression length and analogy to unforced speech.Emotion detection is applied to find an ideal predictable set and to examine the relation of disparate of emotions and speech of the human-being.Speech emotion recognition is a laborious piece of work because human emotion is intellectual, which means it is very difficult to differentiate, also it refers to practical and unrestricted conditions of data addition,indoor and outdoor framework,environmental noise, radar and intention problems;all together human emotion can only be checked in assorted particular moments throughout a long monitoring process; speech data with emotional classify is usually limited.**

INTRODUCTION

Machine identification of emotional content in speech is critical in many human based systems, such as behavioral health observation and empathetic conversational systems[5].

Emotions partake in human communications and effectively monitor the emotion states[1]. In the last few years, human-computer interactions have become representative of pragmatic communal interactions[2]. Now AI coordinators have the ability to recognize human speech. People also understand information through emotional signals, and current AI technologies are inadequate to obtain emotional communication.Because of this gap and benefits come to light, intelligence has led to attracting humans to speech emotion recognition[3].

Speech Emotion Recognition (SER) can be defined as uprooting of the emotions of the person from his or her speech[4].

To communicate productively with people, the process needs to recognize the emotions in speech. Therefore, It is important to grow machines that can acknowledge paralinguistic information like emotion to have effective clear communication like humans. One Significant Part in a para linguistic information is Emotion, which is carried along with speech.The purpose is to make the human-machines interaction natural..A lot of machine learning algorithms have been developed and tested in order to categorize these emotions carried by speech.

For Example, in call centers, tracking customers' emotion states can be practical for measurement and the calls from annoyed or rude customers can therefore be assigned to experienced agents[5].

Speech is one of the communication course that emotions could have a serious impact on. Technically, emotions affect both the vocal sound characteristics and linguistic content. In this study, we focus on the change of vocal sound characteristics to acknowledge the underlying emotions in speech of the person.

Auxiliary it is hard to get access to the existing corpora of voluntary emotional speech. As a result, in many cases one of two replacement strategies is followed: performed speech is recorded from the speech of professional actors.

Nowadays deep learning methods have been introduced to this field. In Multilayer perceptron (MLP) was used on the top of traditional utterance-level features and achieved a significant improvement on accuracy compared with Convolution Neural Network (CNN) used MLP to learn the temporary acoustic features, followed by traditional statistical functions to construct utterance-level features[1].

We compared the two datasets of models. For both, we extract three hand-crafted features from the audio signal[4]. In the first dataset model, the extricated features are used to guide the two machine learning datasets i.e SVM and Decision tree , whereas the second approach is based on deep learning wherein we used CNN and MLP for classification.Overall, we show that MLP based models skilled over a few hand-crafted features can achieve performance comparable to the other classifiers for emotion
recognition.

LITERATURE REVIEW

This literature review processes three prime details of SER. The first one is the correct demonstration of an emotional speech database for calculating system production The second is choosing the appropriate features for speech depiction and the third one is analyzing the structure and then finalizing a suitable classifier.Speech emotion recognition components a front-end processing unit that fetches all features from the speech data and then apply them to a classifier for the prediction of the emotion present in the speech.The concept of SER is not new and has been studied and investigated for quite some time , initially HMM models were used but the introduction of deep neural network has drastically showed better results . They are more powerful in modeling random mappings.We have done a comparative analysis between Deep learning models like mlp , cnn that are trained end to end and lighter machine models like decision tree and svm.The results show that MLP get defeat SVM in overall emotion classification performance. Even though, the training for SVM was more quickly when compared to MLP , the ultimate precision of MLP was more than SVM Different span of CNN are being executed and trained ,emotion is deduced from speech wave using filter banks.As a result it was finalized that to train a cnn model huge amount of data set is compulsory MLP shows a effective evaluation than Decision tree so the classifiers used is our project is MultiLayer

Perceptrons Classifier (MLP) having an ability of getting right answer of an XOR operator and also many different non -linear functions and has shown an accuracy of 81.51 % and is able to identify 8 emotions.We also further analyze the details occupied in various modalities and how their permutation have an impact on the conduction.

Proposed System

For the system, first we have selected the RAVDESS and TESS dataset for classification. We have combined both the datasets into a single set after that we have extracted the features i.e. MFCC, MEL SPECTROGRAM and Chroma from the set and we have splitted the data into train dataset (75%)
and test dataset (25%).We have trained the model on training dataset using MLP(Multi-layer perceptron) with activation function relu and tested the model on test dataset and got the accuracy of 81%.Saving the model in the pickle file we have connected it to flask framework. Using HTML, CSS, MySQL and flask framework we have developed a fully working website for emotion detection. In the website first you have to register/login after that you can upload an audio file and the emotion will be detected automatically and will be displayed on the screen. User can also view the history of the audio files and emotions which were uploaded earlier by the same user.We have developed a flexible and user-friendly website for speech emotion recognition.

Methodology

1]Dataset

In this work, we have used two datasets i.e.RAVDESS and TESS.RAVDESS stands for Ryerson Audio-Visual Database of Emotional Speech and Song and TESS stands for Toronto emotional speech set .The format of the audio file is a WAV format for both the datasets.

RAVDESS

- It is a dataset of emotional speech which contains 7356 files. [11]
- It is a gender balanced database which consists of 24 actors, speaking lexically-matched statements in North American accent.
- It includes calm, happy, sad, angry, fearful, calm, disgust and surprise.

TESS

- In this dataset two actresses (aged 26 and 64 years) have spoken set of 200 target words in the carrier phrase "Say the word _' and it shows each of seven emotions (anger, fear, pleasant surprise, happiness, sadness, disgust and neutral)
- It has 2800 data points (audio files) in total.
- It consists 200 targeted words in audio file and dataset is organised in the manner that both female actors and their emotions are contained within its own folder.[12]

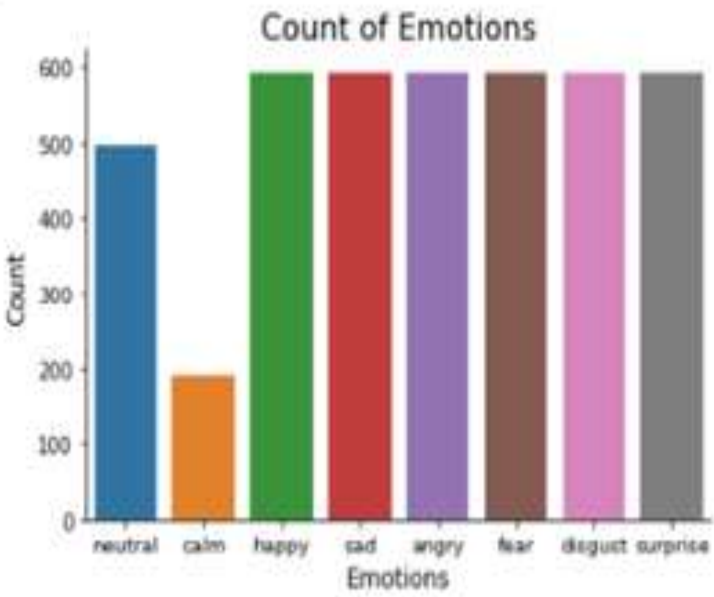


Fig 1

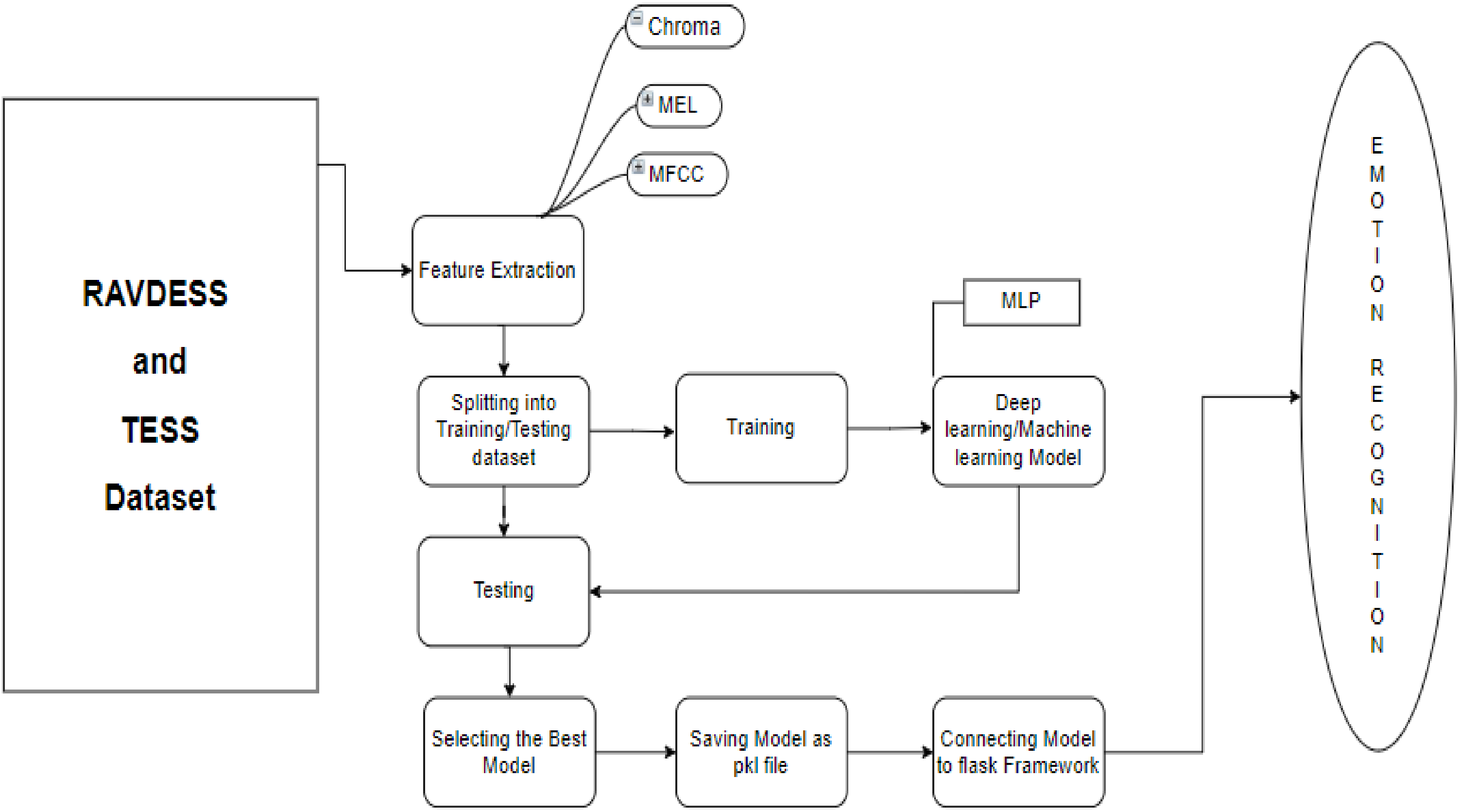


Fig 2

2]Feature Extraction

Features used for training are:

*MFCC:*Mel-frequency cepstral coefficients (MFCCs) are coefficients together from MFC. The power spectrum of sound is represented by MFC mel-frequency cepstrum)**.**We can easily derive MFCC's from cepstral representation of the audio clip (a nonlinear "spectrum-of-a-spectrum"). The key difference between the cepstrum and the mel-frequency cepstrum is that in normal cepstrum the frequency bands are linearly-spaced on the mel scale, while in mel-frequency cepstrum the frequency bands are equally spaced on the mel scale, which resembles the human auditory system's response. This type of frequency warping can be used for better depiction of sound.

MFCCs can be derived as follows:

1. Take the Fourier transform of an audio signal.
2. Using triangular overlapping windows calculate the powers of the spectrum acquired from the mel scale.
3. Take the logs of the powers that has been computed in step 2, at each of the mel frequencies.
4. Consider it as a signal and compute the cosine transform of the list of powers of mel log.
5. The amplitudes of the derived spectrum are the required MFCCs.

This process can be performed in different ways, for example: for mapping the scale one can use different shape or different spacing of the windows or even one can add other features like "delta" and "delta-delta" coefficients.

MELSPECTROGRAM: A mel spectrogram is a spectrogram in which the frequencies of audio files are converted to mel scale[13]. Its object represents an acoustic time-frequency representation of a sound i.e the power spectral density $P(f, t)$.Number of points are sampled around equally spaced times t_i and frequencies f_j (on a Mel frequency scale).

The mel frequency scale is defined as:

$$\text{mel} = 2595 * \log 10\ (1 + \text{hertz} / 700),$$

and its inverse is:

$$\text{hertz} = 700 * (10.0^{\text{mel} / 2595.0} - 1).$$

*Chroma:*Chromagram or chroma feature is associated with twelve different pitch classes. Chroma-based features are stated as "pitch class profiles" and these pitch class profiles are the primary tool for interpreting the music whose pitches can be classified generally into twelve categories. The most significant property of chroma features is that they catch the harmonic and melodic attributes of music and also they are reluctant to change in timbre and instrumentation. We have used chroma_stft which evaluates chromagram from a waveform or power spectrogram.

3]Algorithms

The various models which we have used in our project are: MLP, CNN, Decision tree and SVM.

Machine Learning: In machine learning, the computer automatically extracts the algorithm from the given data, i.e. machine studies the behaviour of data.

a) *SVM:* It does classification by finding a hyperplane with the largest distance margin between the given classes . It transforms the data into multiple dimensions by using a kernel function. Large feature space can be handled using SVM. It has less error rate than other algorithms. Here we use a soft margin approach to reduce overfitting. The algorithm is not effective if the dataset has some noise.

b) *Decision tree:* Here we use tree representation to classify data. Selecting an important attribute is the first step towards this algorithm. At each node we have to decide which is the most important attribute on the basis of which next split can be done. Concepts like information gain, gain ratio method, gini index are used to check which is the most important attribute or the splitting attribute. For example, Information gain can be calculated as the entropy of a particular attribute subtracted from the entropy of the entire dataset. But one of the drawbacks in the information gain method is that it is more biased towards attributes with a large number of values.Gain ratio method is basically normalization of information gain which can help overcome the drawback of information gain method.

Deep Learning : In this type models used are of multiple layers and they extract features which are of high level . The word deep describes the depth of the number of layers used to process the raw input.

a) *CNN :* It is mainly used for the purpose of classification of images and recognition of voice. Features are extracted from the input through a convolutional layer and passed to the pooling layer to reduce the size of the feature map and a fully connected layer connects the neurons of both the layers and then activation function is used to learn the approximation.

b) *MLP:* These are the models that are fully connected in which all of the neurons of every layer are connected to all neurons of the next layer. It has 3 layers (input, hidden and output) and is capable of handling non linear functions

4]COMPARATIVE ANALYSIS

Classifier	Accuracy
Decision Tree	74.15
Support Vector Machine	78.58
Convolution Neural Network	80.28
Multilayer perceptron	81.51

5]IMPLEMENTATION AND RESULTS:

Here we have used libraries like librosa, sklearn.

Libraries:

Librosa: It is a library found in python. The use of this library is analysis of speech, audio and voice files. It consists of a load function which is used to load the audio file by specifying the path of the audio file. This function returns two parameters: the time and audio file's sampling rate.

Sklearn: it is a library which provides machine learning algorithms and tools which can be used in preprocessing the data, fitting a model,evaluation of the model. Here in our project we have imported neural_network, metrics and used algorithms like MLP,SVM,decision tree,CNN and tools like confusion matrix.

We have divided our dataset into 2 parts 75 % for training and rest 25 % for testing of data.

here emotion representation is as follows:

emotions={
  0:neutral,
  1:calm,
  2:happy,
  3:sad,
  4:angry,
  5:fearful,
  6:disgust,
  7:surprised
}

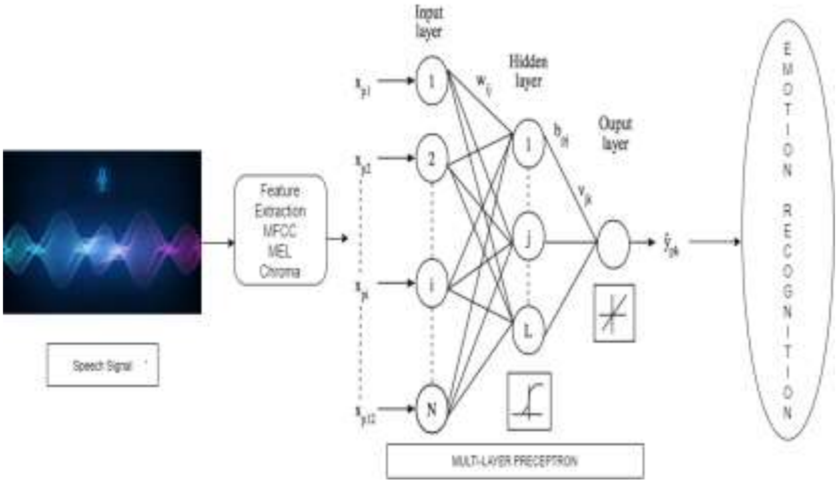
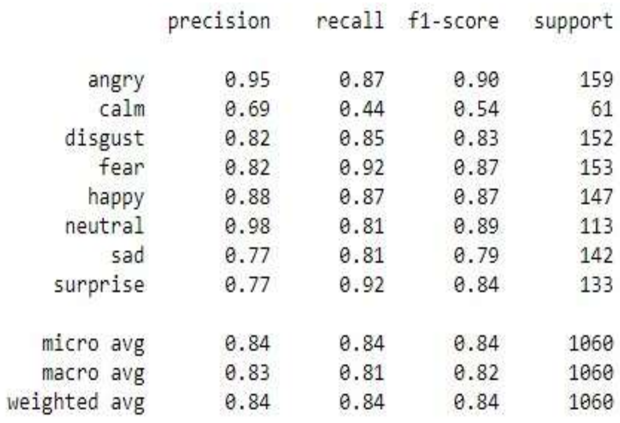


Fig 3

Evaluation using MLP

Formula for Precision =tp/tp + fp

Formula for Recall: =tp/tp + fn



```
              precision    recall  f1-score   support

       angry       0.95      0.87      0.90       159
        calm       0.69      0.44      0.54        61
     disgust       0.82      0.85      0.83       152
        fear       0.82      0.92      0.87       153
       happy       0.88      0.87      0.87       147
     neutral       0.98      0.81      0.89       113
         sad       0.77      0.81      0.79       142
    surprise       0.77      0.92      0.84       133

   micro avg       0.84      0.84      0.84      1060
   macro avg       0.83      0.81      0.82      1060
weighted avg       0.84      0.84      0.84      1060
```

Fig 4

Confusion matrix

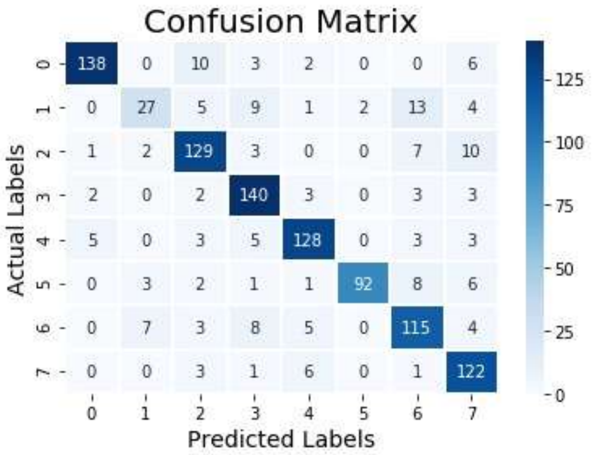


Fig 5

CONCLUSION AND FUTURE WORK:

In this task we have done a comparative study between lighter machine models like decision tree and SVM deep learning models like cnn and mlp, and have also discussed the features that play an important role in reaching a high accuracy in our model. It has been concluded that deep learning models provide high accuracy as svm could be trained faster but the accuracy of 81.51 % given by mlp is much more and could identify eight emotions. It can also be inferred that CNN models require a large amount of data set to be trained. We have extracted three key features from the audio file.Different feature and algorithm combinations give different accuracies for our dataset. We have tried to extract one more feature from audio; LFCC.But the accuracy of our model decreased as compared to earlier when we had only used three features. We got less accuracy from all the four algorithms. The accuracy of our project can be improved by using more features like zero_crossing_rate, chroma_cens, Spectral Rolloff, Spectral Centroid, Pitch, Spectral entropy, Spectral flux, Energy, etc. To achieve better accuracy we can double the size of the dataset. We can also investigate the effect of noise in audio files in our accuracies of the model, that is we can record our own audio with some noise in the background with a particular emotion and check for accuracy of the emotion detected.

REFERENCES

1) Anagnostopoulos, C.-N., Iliou, T. and Giannoukos, I. Features and classifiers for emotion recognition from speech: a survey from 2000 to 2011. Artificial Intelligence Review 43, 2 (2015), 155177.
2) https://link.springer.com/chapter/10.1007/978-3-540-24842-2_3.
3) T. L. Nwe, S. W. Foo, and L. C. De Silva, "Speech emotion recognition using hidden Markov models".
4) https://www.researchgate.net/publication/299185942_Human_speech_emotion_recognition#:~:text=Speech%20Emotion%20Recognition%20(SER)%20can,his%20or%20her%20speech%20signal.&text=The%20potential%20features%20are%20extracted,between%20emotions%20and%20speech%20patterns.
5) Mao, Q., Dong, M., Huang, Z. and Zhan, Y. Learning salient features for speech emotion recognition using convolutional neural networks.
6) Dellaert, F., Polzin, T. and Waibel, A. Recognizing emotion in speech. In Proceedings of ICSLP 3, (Philadelphia, PA, 1996).
7) Gaurav Sahu "   Multimodal Speech Emotion Recognition and Ambiguity Resolution " University of Waterloo ,Ontario, Canada.
8) K.S. Rao, T.P. Kumar, K. Anusha, B. Leela, I. Bhavana, Gowtham S.V.S.K. "Emotion Recognition from Speech" (IJCSIT), 2012
9) https://medium.com/nybles/a-brief-guide-to-convolutional-neural-network-cnn-642f47e88ed4
10) https://medium.com/@ODSC/the-complete-guide-to-decision-trees-part-1-aa68b34f476d
11) https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0196391
12) https://tspace.library.utoronto.ca/handle/1807/24487
13) https://medium.com/analytics-vidhya/understanding-the-mel-spectrogram-fca2afa2ce53#:~:text=A%20mel%20spectrogram%20is%20a,converted%20to%20the%20mel%20scale.&text=What's%20amazing%20is%20that%20after,a%20couple%20lines%20of%20code.