Sentiment Recognition Through Speech: Comparing SVC model and MLPClassifier

Rishi Jain
SET Department
Asian Institute of Technology
Bangkok, Thailand
st122603@ait.asia

Shubhangini Gontia
SET Department
Asian Institute of Technology
Bangkok, Thailand
st121473@ait.asia

***Abstract*—Emotion recognition through speech focuses on categorize the audio files to specific emotion. The audio files are used to read the frequency and the pitch. To achieve the goal to recognize the basic emotions like calm, happy, sad and angry, we are comparing the accuracy for the two models, MLPClassifier model and SVC model. The dataset we used is RAVDESS and used MFCC, chroma and mel for the audio signals.**

Keywords—emotion recognition, mlpclassifier, svc, mfcc, mel feature.

I. Introduction

In naturalistic human-computer interaction (HCI), speech emotion recognition (SER) is becoming increasingly important in various applications.[3] Speech Emotion Recognition can be defined as an act of attempting to recognize human emotion and affective states from speech. Voice often helps in understanding the underlying emotions of a person through his/her tone and pitch. Even animals like dogs and horses are able to understand human emotion through their tone and pitch. SER is tough because emotions are subjective and annotating audio is challenging. This paper compares the two model Multi- Layer Perceptron (MLP) and SVC model. SER can be used for various other speech processing applications. The developed model can be used in call centers where the employees could understand the pitch and tone of the customers and react accordingly. This can help the call centres employees predict customers' emotions from speech and can improve their service. This in turn helps them in converting more people and providing customer satisfaction.

II. Related Work

A. *Emotion Recognition from Speech using Spectrograms and Shallow Neural Networks*

SER (Speech Emotion Recognition) system was proposed in which the authors combined the power of DL models in self pattern recognition together with the ability of working on small databases(RAVDESS, RML, EMOD). Spectrograms were generated for audio file and 1d cnn was applied for classification.[2]

B. *Speech Emotion Recognition using Neural Network and MLP Classifier*

This paper presents the use of Neural Networks to classify the emotions from a given speech, known as Speech Emotion Recognition (SER). Speech Emotion Recognition helps to classify elicit specific types of emotions. In the experiment they used MLP and Neural Network to get an accuracy of 70.28% using logistic activation function and 65% using ReLU function for data processing. Testing the with an input audio sample also gave them same results.[4]

III. Proposed Methodology

The underlying emotion in our speech is reflected in our voice through tone and pitch. In this paper we aim to classify basic types of emotions such as sad, happy, neutral, angry, disgust, surprised, fearful and calm. In this paper the emotions in the speech are predicted using neural networks. Multi-Layer Perceptron Classifier (MLP Classifier) and Support Vector Classification(SVC) is used for the classification of emotions. RAVDESS (Ryerson Audio-Visual Database of Emotional Speech and Song dataset) is the dataset used in this paper.

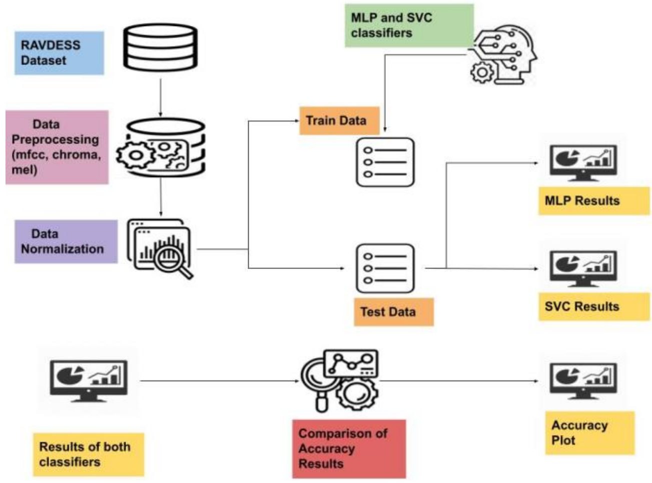


Fig. 1. Study Methodology

A. *RAVDESS Dataset*

RAVDESS dataset has recordings of 24 actors. The emotions contained and labelled in the dataset are as 01-'neutral', 02-'calm', 03-'happy', 04-'sad', 05-'angry', 06-'fearful', 07-'disgust', 08 -'surprised'. The dataset contains all expressions in three formats, which are: Only Audio, Audio-Video and Only Video. Since our focus is on recognize emotions from speech, our model is trained on Audio-only data. Two fixed statements are vocalized by all the 24 actors for all the 8 emotions, with each statement repeated twice. All emotional expressions are uttered at two levels of intensity: normal and strong, except for the 'neutral' emotion, it is produced only in normal intensity. Thus, the portion of the RAVDESS, that we use contains 60 trials for each of the 24 actors, thus making it 1440 files in total. The dataset is labelled in accordance with the decimal encoding. Ever file has a unique filename. The filename is made up of 7-part numerical identifier, the 3rd numerical part of the filename denotes a label to the corresponding emotion.[4]

B. *Feature Extraction(mfcc,chroma and mel)*

- MFCC: Mel Frequency Cepstral Coefficients (MFCC) feature extraction technique basically includes windowing the signal, applying the DFT, taking the log of the magnitude, and then warping the frequencies on a Mel scale, followed by applying the inverse Discrete Cosine Transform (DCT).

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

- Chroma: Compute a chromagram from a waveform or power spectrogram. .The chroma is figured by including the log-repeat size range across octaves. The coming about plan of chroma vectors is known as chroma-gram.

$$Cf(b)= \Sigma Z-1z=0|Xlf(b+z\beta)|| \qquad (2)$$

- Mel: Compute a mel-scaled spectrogram. If a spectrogram input S is provided, then it is mapped directly onto the mel basis by mel_f.dot(S). If a time-series input y, sr is provided, then its magnitude spectrogram S is first computed, and then mapped onto the mel scale by mel_f.dot(S**power).

$$M(f)=1125\ Ln(1+f/700) \qquad (3)$$

C. *Multi-layer Perceptron Classifier*

Multi-layer Perceptron (MLP) classifier, this model optimizes the log-loss function using LBFGS or stochastic gradient descent. MLP Classifier implements a Multi-Layer Perceptron (MLP) algorithm and trains the Neural Network using Backpropagation.

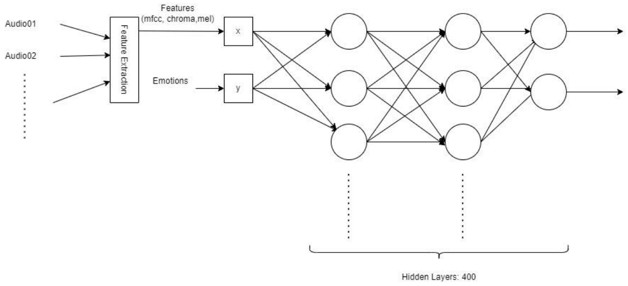


Fig. 2. Multi-layer Perceptron Classifier

As shown in the Fig.1. Once the features are extracted from the audio file, namely mfcc, chroma and mel, then the three outputs are concatenated to one list of lists with all the frequency values. The emotions list will contain the observed emotion and in our analysis we observe three emotion calm, angry and happy. These lists of features and emotions are the inputs for the MLP classifier and will be trained under the hidden layers. The hidden layer uses an activation function to act upon the input data and to process the data. For the better outcome we have used 400 hidden layers. After the model is trained the output is used to predict the accuracy for the trained model.

D. *Support Vector Classification*

Support Vector Classification(SVC) implementation is based on libsvm. The multiclass support is handled according to a one-vs-one scheme. Support Vector Machine is efficient in predicting emotions for sound input with no discrepancy, in presence of noisy input it deviates from its prediction. SVM only classifies using a single plane and restricts the prediction. The results show that the system using the Support Vector Machine i.e., SVM has a more computational time, even though having a decent accuracy. As SVM only works on a single plane and therefore it faces problems addressing complex time series based data. [1] The Linear Support Vector Classifier (SVC) method applies a linear kernel function to perform classification and it performs well with a large number of samples. If we compare it with the SVC model, the Linear SVC has additional parameters such as penalty normalization which applies 'L1' or 'L2' and loss function.

For our analysis, we used the parameters kernel as ‘rbf’ which specifies the kernel type to be used in the algorithm, gamma as scale which is kernel coefficient for ‘rbf’, maximum iteration as 500 to get the best accuracy.

IV. Impelementation

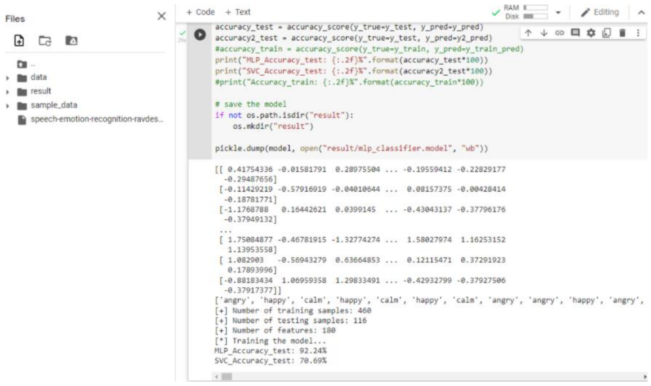
First, we will analyse the predicted accuracy for MLP classifier. The extracted features will be split into train and test data with the ratio is 80:20. 80% train data will be used to train the MLP model and ret 20% will be used to predict the accuracy for the model. Initializing the classifier, set the model parameters ‘alpha’ as 0.05, the batch size is set to 256, epsilon 1e-08, hidden layer as 400, learning rate as ‘adaptive’ keeps the learning rate constant to ‘learning_rate_init’ as long as training loss keeps decreasing. Each time two consecutive epochs fail to decrease training loss by at least tol, or fail to increase validation score by at least tol if ‘early_stopping’ is on, the current learning rate is divided by 5 and the number of maximum iterations as 500. Once the MLP is initialized, train the model with input feature train data and emotion train data.

After the model is trained, find the accuracy for the predicted values, by input of validation data. Store the accuracy to plot the graph for the comparison between the accuracy of MLPClassifier and SVC model.

Now, Initialize the SVC model with the model parameters kernel as ‘rbf’ which specifies the kernel type to be used in the algorithm, gamma as scale which is kernel coefficient for ‘rbf’, maximum iteration as 500 to get the best accuracy.

Once, the SVC model is initialized, train the model with training data of features and emotions. After training the model, find the accuracy for the predicted values by input of validation data of features. Store the accuracy for the comparison with the MLP model.

Following are the results for the accuracies obtained for the MLPClassifier and SVC model with training sample of 460 and validation sample of 116 with 180 number of features.



```
accuracy_test = accuracy_score(y_true=y_test, y_pred=y_pred)
accuracy2_test = accuracy_score(y_true=y_test, y_pred=y2_pred)
#accuracy_train = accuracy_score(y_true=y_train, y_pred=y_train_pred)
print("MLP_Accuracy_test: {:.2f}%".format(accuracy_test*100))
print("SVC_Accuracy_test: {:.2f}%".format(accuracy2_test*100))
#print("Accuracy_train: {:.2f}%".format(accuracy_train*100))

# save the model
if not os.path.isdir("result"):
    os.mkdir("result")

pickle.dump(model, open("result/mlp_classifier.model", "wb"))
```

```
[[ 0.41754336 -0.01581791  0.28975504 ... -0.19559412 -0.22829177
  -0.29487656]
 [-0.11429219 -0.57916919 -0.04010644 ...  0.08157375 -0.00428414
  -0.18781771]
 [-1.1768788   0.16442621  0.0399145  ... -0.43043137 -0.37796176
  -0.37949132]
 ...
 [ 1.75084877 -0.46781915 -1.32774274 ...  1.58027974  1.16253152
   1.13953558]
 [ 1.082903   -0.56943279  0.63664853 ...  0.12115471  0.37291923
   0.17893996]
 [-0.88183434  1.06959358  1.29833491 ... -0.42932799 -0.37927506
  -0.37917377]]
['angry', 'happy', 'calm', 'happy', 'calm', 'happy', 'calm', 'angry', 'angry', 'happy', 'angry',
[+] Number of training samples: 460
[+] Number of testing samples: 116
[+] Number of features: 180
[*] Training the model...
MLP_Accuracy_test: 92.24%
SVC_Accuracy_test: 70.69%
```

Fig. 3. Accuracy for MLPClassifier and SVC

For MLPClassifier we get 92.34% accuracy whereas we get 70.69% for SVC model for RAVDESS dataset with emotions calm, angry and happy.

Following is the graphical representation of the obtained accuracies for both of the model.

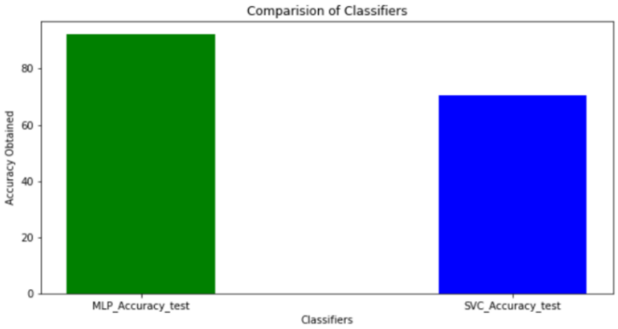


Fig. 4. Accuracy graph representation for MLPClassifier and SVC model

V. Conclusion

This paper shows that MLPClassifier is the better for the sentiment recognition than SVC model. The accuracy for calm, angry and happy sentiments for MLP is 92.43%. This is possible because of the good data preprocessing, where mfcc, chroma and mel where used for feature extraction.

References

[1] Roy, T., Marwala, T., & Chakraverty, S. (2020). Speech Emotion Recognition Using Neural Network and Wavelet Features. Lecture Notes in Mechanical Engineering, 10(4), 427–438. https://doi.org/10.1007/978-981-15-0287-3_30

[2] Slimi, A., Hamroun, M., Zrigui, M., & Nicolas, H. (2020). Emotion recognition from speech using spectrograms and shallow neural networks.

[3] B. R. Sanjita, A. Nipunika, Rohita Desai, "Speech Emotion Recognition using MLP Classifier," in IJESC 2020

[4] Jerry Joy,Aparna Kannan, Shreya Ram and S.Rama "Speech Emotion Recognition using Neural Network and MLP Classifier," IJESC 2020.