e-ISSN: 2582-5208

International Research Journal of Modernization in Engineering Technology and Science

(Peer-Reviewed, Open Access, Fully Refereed International Journal)

Volume:04/Issue:04/April-2022 Impact Factor- 6.752 www.irjmets.com

SPEECH EMOTION RECOGNITION SYSTEM USING ML

Anish Chandra Jojula[*1], **Sayam Bothra**[*2]

[*1,2]Computer Science & Engineering, SRM Institute And Science And Technology, India.

ABSTRACT

Emotions add significance to individual discussions and help us understand each other better. Human-computer interactions have also advanced significantly in recent years in terms of addressing client wants and responsibilities. From this perspective, it would be ideal for robots to understand human emotions automatically in order to increase communication and interaction between the two parties. Some of the research in this field relies on handcrafted functions, while others rely on deep learning (DL) models. This paper presents a SER (Voice Emotion Recognition) system that combines the capability of DL models with the ability to recognize self-patterns.

Keywords: Speech Emotion, Emotion, Emotion Recognition, MLP Classifier , SVM Algorithm.

I. INTRODUCTION

As listeners, we react to the speaker's emotional state and adjust our behavior based on the sort of emotion expressed by the speaker. Recent technological advancements have enabled people to interact with computers through non-traditional means (e.g., speech, gesture, and facial expression) such as voice, gesture, and facial expression. This interaction is still missing emotional elements. It has been improved that in order to attain a human emotional computer, smart interaction is required. In order to communicate with users in a natural manner, comparable to how individuals interact. Many of the research have included ordinary individuals in the human-computer interaction and interaction. Before absorbing semantic information in his mother's voice, a newborn learns to distinguish emotions. We show some fundamental studies in the field of speech-based emotional recognition. First, we'll give you a quick rundown of recent research in the domains. Following that, we'll go through a method for detecting and classifying human emotions in speech that employs a set of rules. Anger, happiness, fear, sorrow, and neutral are the basic year emotions.

II. LITERATURE REVIEW

1. Speech based Emotion Recognition using Machine Learning

Authors: Girija Deshmukh, Apurva Gaonkar, Gauri Golwalkar, Sukanya Kulkarni

Objectives: A system in which they obtained audio samples of Short-Term Energy (STE), Pitch, and MFCC coefficients in frustration, happiness, and sadness of emotions.

Findings and Conclusions: The whole Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) dataset is manually split into train and test sets. The multi-class Support vector machine (SVM) takes feature vectors as input, which is turned up as a model corresponding to each emotion.

2. A Speech Emotion Recognition Model Based on Multi-Level Local Binary and Local Ternary Patterns

Authors: Asaf Varol

Objectives: SER uses role extraction techniques such as acoustic analysis and analysis of spectrogram to perform SER.

Findings and Conclusions: Results are drawn from experiments using the speech signal spectrogram and Artificial Neural Networks (ANNs).

III. PROPOSED METHODOLOGY

A machine learning (ML) model is used to create the voice emotion detection system. The stages are similar to any other machine learning project, with extra fine-tuning techniques to improve the model's performance. Flowcharts depict the process in a visual manner (see Figure 1). The first step is to gather data, which is crucial, and use the RAVDESS dataset as input. The model in development will learn from the data that is provided to it, and data will drive all of the decisions and results that the model produces. The feature extraction step is a set of machine learning tool stacks that are applied to the obtained data, such as MFCC and chroma.

MFCC:

The MFCC (Mel-Frequency Cepstral Coefficient) is a tool that helps assess a person's vocal tract shape. The vocal tract is represented by the envelope of the temporal power spectrum of the speech signal, which is correctly represented by MFCC.

Chroma:

Chroma-based characteristics, also known as "pitch layer profiles," are a strong technique for evaluating music with usefully grouped pitches (typically twelve categories) and tuning that approximates the scale. One of the most essential qualities of chroma features would be that they capture the harmonic and tonal properties of sound while also being resistant to timbre and instrumentation changes. The following are two of the most important chroma characteristics:

(a) Chroma vector: A 12-element spectral energy representation in which the bins reflect the 12 equal-tempered pitch classes of western-style music (semitone spacing).

(b) The standard deviation of the 12 chroma coefficients is referred to as the chroma deviation.

These methods deal with various data formats and data quality challenges. The third phase, when an algorithm based on the established model is constructed, is frequently considered the core of an ML project. This model learns more about data using machine learning methods, and it responds to any new data it encounters. The final stage is to assess the created model's performance. Developers frequently repeat the stages of creating a model and assessing it in order to compare the performance of various methods. The comparative findings aid in selecting the best suited ML method for the task.

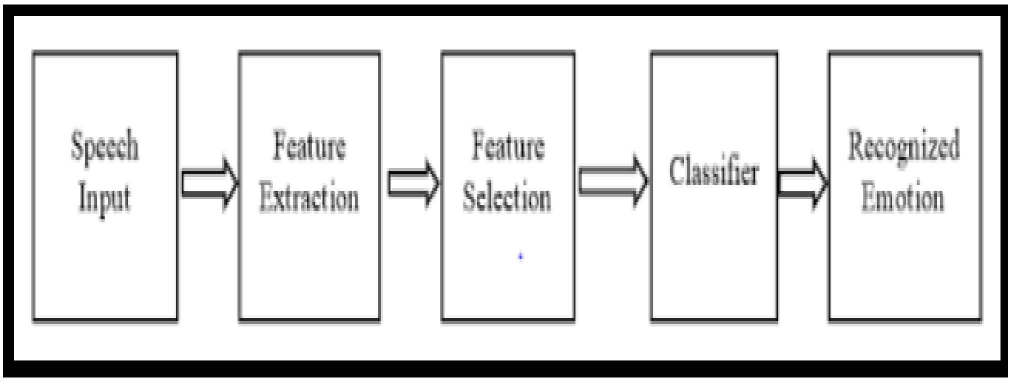


Fig 1: Block Diagram

MLP CLASSIFIER:

They may approximate the XOR operator, as well as many other non-linear functions, according to subsequent work on the multilayer perceptron. Multilayer Perceptron is a type of supervised learning algorithm. They practice How to model input-output pairs' sets and relationships (or dependencies between these inputs and outputs). As a result, the network is merely viewed as a source of input and output. Model parameters without weights and thresholds (offsets). Include an indicator of the number of hidden layers and the number of units in this layer in critical MLP design questions. the total number of hidden units It's unclear how to use it. This is a fantastic place to start. Half the amount of units, one hidden layer The total number of input and output units is the sum.

The Classifier recognizes several categories in the datasets and categories them into various emotions. The model will now be able to recognize the ranges of speech parameter values that correspond to different emotions. If we provide the model an unknown test dataset as an input, it will extract the parameters and forecast the emotion based on the values in the training dataset. The system's accuracy is shown as a percentage, which is the end outcome of our project.

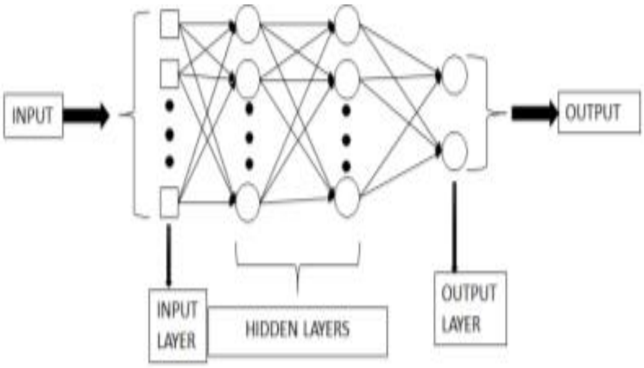


Fig 2: MLP Classifier

SVM:

SVMs (support vector machines) are supervised machine learning techniques that may be used for both classification and regression. However, they are most commonly utilised in categorization difficulties. SVMs were initially presented in the 1960s, but they were enhanced around 1990. In comparison to other machine learning algorithms, SVMs feature a unique implementation method. They've recently gained a lot of traction due to their capacity to handle both continuous and categorical variables.

The dataset's support vectors are unique data points. They are responsible for the hyperplane's creation and are the hyperplane's nearest points. The hyperplane's location would be changed if these points were deleted. The hyperplane is surrounded by decision limits. The support vectors aid in the reduction and expansion of boundary sizes. They are the most important parts of an SVM.

We use the SVM technique to find the spots in both groups that are closest to the road. Support vectors are the names given to these locations. In SVM, the decision boundaries are the two lines that run parallel to the hyperplane. The margin is the distance between the two light-toned lines. When the margin is narrow, the optimum hyperplane shape is found.

In the diagram above, the hyperplane is the middle line. Because the dimension is two-dimensional, the hyperplane is a line in this example. The hyperplane would have been a 2-D plane if we had a 3-D aircraft. The study of the hyperplane necessitates a significant amount of mathematics. We'll take a look at it. However, in order to comprehend a hyperplane, we must first see it. Assume there is a focal point (a blank piece of paper). Imagine a line from the centre cutting across it.

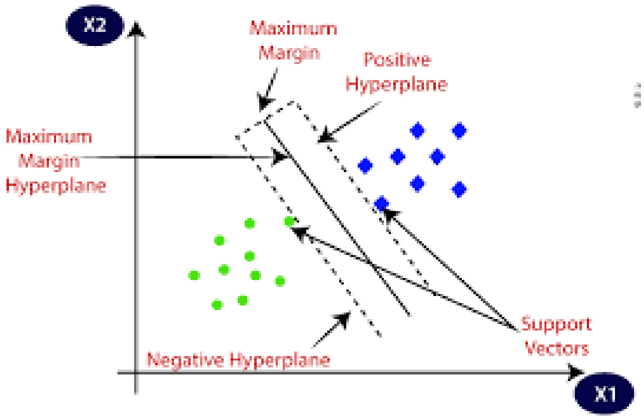


Fig 3: SVM

IV. IMPLEMENTATION

1. DATA COLLECTION

The initial stage in creating the Speech Emotion Recognition system is to gather audio samples that may be used to train the model under various emotional categories. The audio samples are generally wav or mp3 files that may be downloaded for free. The stages that follow are outlined with relation to the TESS dataset trials.

PYTHON LIBRARY

Following data gathering, the next step was to numerically represent these audio recordings in order to undertake additional analysis. This stage is known as feature extraction, and it involves obtaining quantitative values for various audio aspects. Librosa is a Python tool that analyses music and audio. It contains the components required to construct music information retrieval systems. Python's library Librosa is an open Python library that focuses on feature extraction difficulties and provides a wide variety of audio-related features. The library depends on several other libraries which are:

•Numpy• Matplotlib

•Scipy •Sklearn

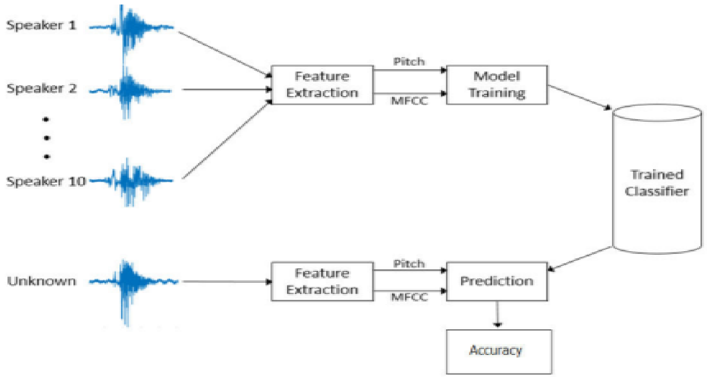


Fig 4: Implementation Process.

FEATURE EXTRACTION:

The process of modifying, reducing, or building features for a dataset is known as feature engineering. Each feature contains numerous values for each frame of the audio stream, as indicated before in the raw data. The frame size and frame overlap values can be modified using frame blocking and windowing techniques to produce precise audio signal values. The average values of distinct attributes for audio signals are also produced using the averaging approach. Each audio signal is now represented by 34 discrete values in the converted data. The choice to reduce the amount of features is critical. The decision to eliminate features is usually based on topic expertise, which might have an impact on the system's performance. Following that, a series of tests are carried out utilizing

FEATURE SELECTION:

The feature extraction procedure is crucial for speech emotion identification. The accuracy of classification results is directly influenced by the quality of the features. In most cases, the feature extraction approach creates handcrafted features based on speech acoustic properties.

METHODS USED:

- A feed forward artificial neural network called a multilayer perceptron (MLP) creates a set of outputs from a collection of inputs. Several layers of input nodes are connected as a directed graph between the input and output layers of an MLP. Back propagation is used by MLP to train the network. MLP is a technique of deep learning. Back propagation is a supervised learning approach used by MLPs to assist them offer the closest emotion using spectrograms.
- Support vector machines are a collection of supervised learning algorithms for classification, regression, and identification of outliers. SVMs vary from other classification algorithms in that they choose a decision boundary that optimises the distance between all classes' closest data points. The maximum margin classifier or maximum margin hyper plane is the decision boundary established by SVMs.

COMPARATIVE ALNALYSIS FOR SPEECH EMOTIONS RECOGNITON USING DIFFERENT MACHINE LEARNING TECHNIQUES:

Performance analysis based on different machine learning techniques for different languages. The comparison shows that the different machine learning methods have been used for recognizing speech emotions system for numerous languages. Based on that, the accuracy has been computed for the best case. Our model is based on negative emotions or neutral emotions like sad, angry, calm and fearful.

V. RESULTS

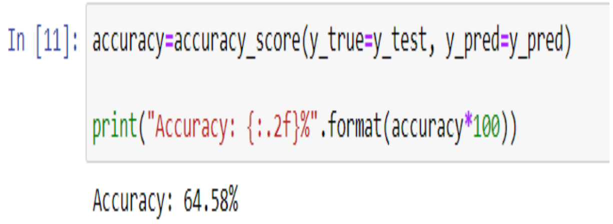
Accuracy was calculated for one emotion at a time. For the MLP classifier.

# Calculate the accuracy of our model.

accuracy = accuracy_score(y_true=y_test, y_pred=y_pred)

# Print the accuracy

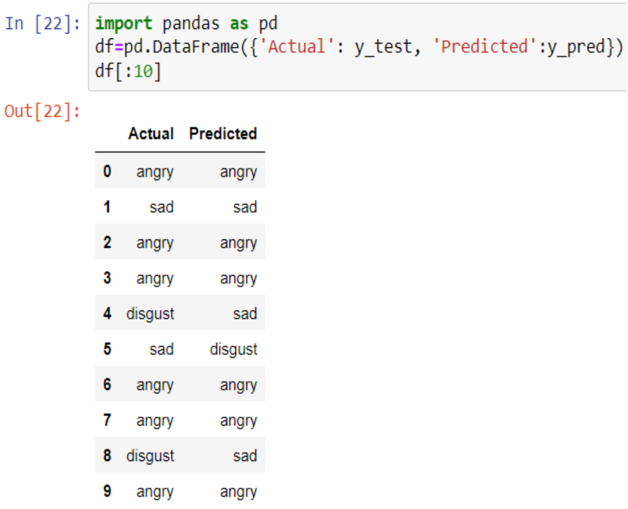
print("Accuracy:{:.2f}%".format(accuracy*100))
Accuracy: 64.58%



```
In [11]: accuracy=accuracy_score(y_true=y_test, y_pred=y_pred)

         print("Accuracy: {:.2f}%".format(accuracy*100))

         Accuracy: 64.58%
```



```
In [22]: import pandas as pd
         df=pd.DataFrame({'Actual': y_test, 'Predicted':y_pred})
         df[:10]
Out[22]:
```

	Actual	Predicted
0	angry	angry
1	sad	sad
2	angry	angry
3	angry	angry
4	disgust	sad
5	sad	disgust
6	angry	angry
7	angry	angry
8	disgust	sad
9	angry	angry

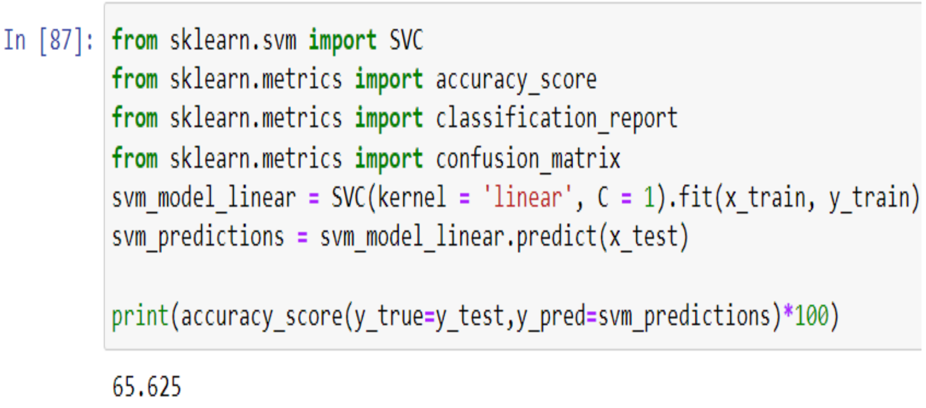
Accuracy was calculated for one emotion at a time. For the SVM.

# Calculate the accuracy of our model.

accuracy = accuracy_score(y_true=y_test, y_pred=y_pred)

# Print the accuracy

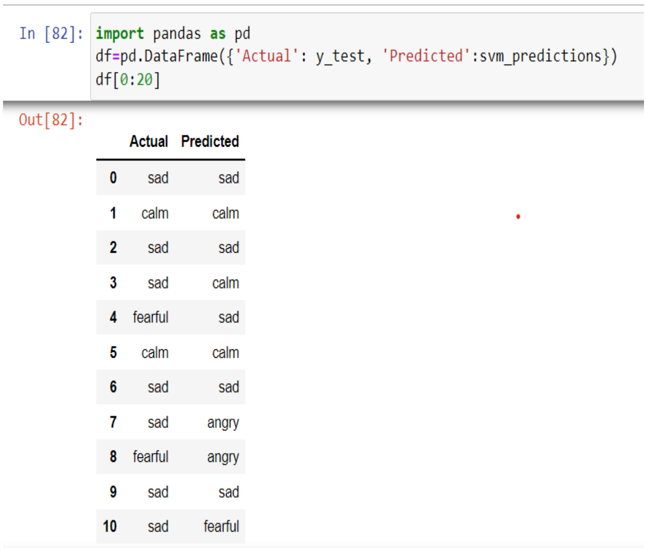
print("Accuracy:{:.2f}%".format(accuracy*100))
Accuracy: 65%



```
In [87]: from sklearn.svm import SVC
         from sklearn.metrics import accuracy_score
         from sklearn.metrics import classification_report
         from sklearn.metrics import confusion_matrix
         svm_model_linear = SVC(kernel = 'linear', C = 1).fit(x_train, y_train)
         svm_predictions = svm_model_linear.predict(x_test)

         print(accuracy_score(y_true=y_test,y_pred=svm_predictions)*100)

         65.625
```



```
In [82]: import pandas as pd
         df=pd.DataFrame({'Actual': y_test, 'Predicted':svm_predictions})
         df[0:20]

Out[82]:
              Actual  Predicted
          0      sad        sad
          1     calm       calm
          2      sad        sad
          3      sad       calm
          4  fearful        sad
          5     calm       calm
          6      sad        sad
          7      sad      angry
          8  fearful      angry
          9      sad        sad
         10      sad    fearful
```

VI. CONCLUSION

This article demonstrates that MLPs are extremely effective in classifying speech signals. A small collection of characters may be easily distinguished even with reduced representations. In comparison to other methodologies, we were able to get better accuracies for individual emotions. The quality of pre-processing has a big impact on a module's performance. Mel Frequency (Mel Frequency) Cepstrum Coefficients are quite reliable. Every human emotion has been meticulously researched, evaluated, and verified for accuracy. The findings show that speech recognition is possible, and that MLPs may be utilised for any task involving speech recognition and proving the correctness of each emotion contained in the speech.

SVM is a complex method that uses kernels to operate as both a linear and non-linear algorithm. There is no shortage of domains and circumstances where SVM may be employed in terms of application areas. It can even be utilised in text categorization because of its ability to cope with high-dimensional spaces. When working with SVM, however, patience is required because tweaking the hyper-parameters and picking the kernel is critical, and the training step takes a long time.

VII. REFERENCES

[1] Rao, K.Sreenivasa, et al. "Emotion recognition from speech." International Journal of Computer Science and Information Technologies 3.2 (2012): 3603-3607.

[2] Sethu, Vidhyasaharan, EliathambyAmbikairajah, and Julien Epps. "Speaker normalisation for speech-based emotion detection." Digital Signal Processing, 2007 15th International Conference on. IEEE, 2007.

[3] Busso, Carlos, et al. "Iterative feature normalization scheme for automatic emotion detection from speech." IEEE transactions on affective computing 4.4 (2013): 386-397.

[4] Schuller, Björn, Gerhard Rigoll, and Manfred Lang. "Speech emotion recognition combining acoustic features and linguistic information in a hybrid support vector machine-belief network architecture." Acoustics, Speech, and Signal Processing, 2004. Proceedings.(ICASSP'04). IEEE International Conference on. Vol. 1. IEEE, 2004.

[5] Z. Yongzhao and C. Peng, “Research and implementation of emotional feature extraction and recognition in speech signal,” Joural of Jiangsu University, vol. 26, no. 1, pp. 72–75, 2005.

[6] L. Zhao, C. Jiang, C. Zou, and Z. Wu, “Study on emotional feature analysis and recognition in speech,” Acta Electronica Sinica, vol. 32, no. 4, pp. 606–609, 2004.

[7] A Speech Emotion Recognition Model Based on Multi-Level Local Binary and Local Ternary Patterns.

[8] Prof. Guruprasad G, Mr. Sarthik Poojary, Ms. Simran Banu, Ms. Azmiya Alam, Mr. Harshith K R “EMOTION RECOGNITION FROM AUDIO USING LIBROSA AND MLP CLASSIFIER” International Research Journal of Engineering and Technology (IRJET), vol8, issue 7, 2021.

[9] Monorama Swain, Aurobinda Routray, Prithviraj Kabisatpathy,”Databases, features and classifiers for speech emotion recognition: a review”, I. J. Speech Technology 2018.

[10] S. G. Koolagudi and K. S. Rao, "Emotion recognition from speech: a review," International Journal of Speech Technology, vol. 15, no. 2, pp. 84–115, 2012.

[11] Awni Hannun, Ann Lee, Qjantong Xu and Ronan Collobert, Sequence to sequence speech recognition with time-depth deperable convolutions.

[12] Lawrence R Rabiner Ronald W Schafer, "Introduction to Digital Speech Processing", Vol. 1, Nos. 1–2 (2007) 1–194, 2007 L. R. Rabiner and R. W... Schafer.

[13] D. Neiberg, K. Elenius, I. Karlsson and K. Laskowski, "Emotion Recognition in Spontaneous Speech,"