Proceedings of the Sixth International Conference on Computing Methodologies and Communication (ICCMC 2022)
IEEE Xplore Part Number: CFP22K25-ART; ISBN: 978-1-6654-1028-1

2022 6th International Conference on Computing Methodologies and Communication (ICCMC) | 978-1-6654-1028-1/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICCMC53470.2022.9753976

Speech Emotion Recognition using Machine Learning

KOTIKALAPUDI VAMSI KRISHNA
B.E - CSE
Sathyabama Institute of Science and Technology, Chennai, India.
vamsikrishnakotikalapudi04@gmail.com

NAVULURI SAINATH
B.E – CSE
Sathyabama Institute of Science and Technology.Chennai, India.
sainadhnavuluri12@gmail.com

A.MARY POSONIA
Associate professor , Dept of CSE
Sathyabama Institute of Science and Technology. Chennai, India.
maryposonia.cse@sathyabama.ac.in

***Abstract* — The aim of the paper is to detect the emotions which are elicited by the speaker while speaking. Emotion Detection has become a essential task these days. The speech which is in fear, anger, joy have higher and wider range in pitch whereas have low range in pitch. Detection of speech is useful in assisting human machine interactions. Here we are using different classification algorithms to recognize the emotions , Support Vector Machine , Multi layer perception, and the audio feature MFCC, MEL, chroma, Tonnetz were used. These models have been trained to recognize these emotions (Calm, neutral, surprise, happy, sad, angry, fearful, disgust). We got an accuracy of 86.5% and testing it with the input audio we get the same.**

Keywords — Detection, Speech Input, Feature Extraction, SVM

I. INTRODUCTION

One of the quickest and most normal ways of communicate is to give indications. Emotions can be of different types and can be Expressed in different ways, Facial Expression is one of the most prominent method for recognizing Emotions, Using Facial Expression we can detect Emotions [1].The Other Way of Communication is through Sign Language[2]. The utilization of sound signs is a quick and effective method for interacting with a human machine. All thoughts are utilized by individuals to all the more likely comprehend the message they are getting. Emotion Detection is a troublesome errand for the machine, however then again, it is typical for people. Along these lines, emotion information is utilized by the ability to understand people on a deeper level framework, which works on the connection between the machine and the individual. The feelings of a lady or a man are communicated through words to decide the feelings of conversation. It is unknown as of now what he will do in the wake of leaving the post. Changes happened because of various paces, examples, expressions, and speakers that influence conversation attributes. Various feelings can be communicated in single word, and there are various grammatical forms for each unique feeling, making it hard to separate these words. Hate speech are impacted by the way of life of the speaker and the climate, which causes another issue in light of the fact that the language differs relying upon the climate and social changes. Temporary feelings are enduring and there are two kinds of feelings, and the sorts of feelings communicated by the beneficiary are unclear. Feelings that are acknowledged by the language of discussion can be independent of the speaker or might be affected by the speaker.[3]we can also detect Emotions from different machine learning algorithms like CNN , KNN , Random Forest.

II. LITERATURE SURVEY

Using Machine learning, detecting emotion of the speaker is possible. There are few references to do it. [4] In this paper, they proposed the use of a Gaussian investigation that can separate conversation feeling level dependent on I-vectors demonstrating the dispersion of MFCC usefulness. A review dependent on the IEMOCAP corpus shows that the foundation of the GPLDA surpasses the foundation of the SVM and is less delicate to the I-vector, so the normal level makes it more powerful to change rules during framework improvement.

[5] To work on the capacity to continually recognize feelings from conversation, we are upholding an instructing blunder-based way to deal with learning and an organization that recursively creates memory. In such manner, with the assistance of two continuous RNN (Recurrent Neural Networks) strategies, the main model is utilized as a computerized code to recover the first substance and the subsequent model is utilized for passionate forecast. RE (Reconstruction-error-based) of the primary resource is utilized as an extra resource, and is matched to the main resource and put in the subsequent class. The possibility of the framework is that the framework can concentrate on its "shortcomings" in RE. A RECOLA database - based review shows that a given framework is better than a base framework without RE information as far as the Concordas coefficient.

[6] Social media correspondence is one of the main parts of compelling correspondence with a Computer. For this sort of association, an unmistakable comprehension of the significance of the word and language arrangement and a familiarity with the feelings contained in the conversation are vital to further develop execution. The language used to communicate feelings passes on Emotions like sadness ,fear, joy, and distress. In the main phase of the feeling forecast framework introduced in this article, various sorts of feelings are recognized. The subsequent advance is to utilize a neural organization to anticipate the following series of feelings. Subsequent to joining the diverse discourse images at each point, the grouping is deducted dependent on the communication characters of one tenth of a second. The issue of forecast is that the neural organization is diverted in a nonlinear course, and the change is characterized as the hypothesis of time data. The best gauge of Random Forest accomplished because of anticipating results utilizing the organization is 86.25%.

[7] Security (Network protection) is a significant issue today and with trend setting innovation. It is critical that network safety is in excess of a mysterious framework to ensure against cybercrime. Equivalent biometric information and action arranged humanities permit applications to get to explicit or general data. The principle motivation behind this article is to investigate feeling based talk utilizing worldview acknowledgment, and affectionate displaying with neighbors. The five signs are cepstrum, mel frekans cepstrum, order, cepstrum. The calculation utilizes pockets to prepare the kNN majority. Reconciliation is done straightforwardly. The outcomes show that the greatest presentation gain is accomplished utilizing two distinct kNNs rather than utilizing one kNN.

[8] The test is to keep up with the force of the Delete-Speech Processing System in the midst of commotion. This page shows a wide data transmission that can work on the responsiveness of the emotion acknowledgment framework when indications are harmed because of chosen clamor. This page shows it as well Settling on decisions dependent on explicit prerequisites can give us the greatest aspect to find the best solution.

III. PROPOSED SYSTEM

The main objective of this project is to detect the emotion of the speaker. Before Speech Emotion Detection was carried out as machine learning (ML). The execution steps are contrasted with other ML undertakings, and better plan processes are improved. The initial step is to accumulate data, which is vital. The model being created depends on the data gave and all choices and reactions to the model being created depend on the information. The subsequent advance, called assembling properties, is to gather countless assignments that are performed on the gathered information. This methodology resolves many issues connected with data and data quality. The third step is frequently viewed as the way in to a ML project that upholds calculation improvement. This model uses ML algorithms to learn data and figure out how to get all significant data. The last advance is to assess the presentation of the introduced model. Looking at the outcomes will assist you with picking the most suitable ML calculation for the issue. In this paper, we showed speech emotion recognition (SER) utilizing AI calculations to find out feelings. The activity of a feeling acknowledgment structure can altogether change the general activity of the structure in numerous ways and give a greater number of advantages than the activity of the program. This review tells the best way to understand intense sounds, how to work on the current structure as far as data, how to choose elements, and how to group emotional sounds dependent on feelings.

A.SPEECH INPUT MODULE

Input to the system is speech. The digital representation of the given sound through the sound file is currently handled.

B. FEATURE SELECTION AND CLASSIFICATION MODULE

In this process we extract features from the input data. After extracting, selection of required features is done. Selected features are converted to csv files. Those files are used for training and testing of the classifier. We train our classifier with training data and then we proceed for testing data. Using testing data we find out the accuracy and classification report of the trained classifier. Emotions can be detected from audio files because audio files contains different parameters . Parameters can change the emotion information[9].Voice frequently returns hidden feeling through pitch and tone. The objective of feature extraction is to get useful feature from audio file for feelings. The Audio files contain a lot of information other than emotion detection. Therefore research on how to extract and which parameters to extract are of great important[10].Features are extracted from the audio file given as information. The features are MFCC, Mel, Chroma, Tonnetz .Emotions can also be recognized by combining the Mel Frequency Cepstral Coefficients (MFCC) with the vibration rate(PITCH) in order to characterize the emotion according to its respective vocal speech signals[11].

C.RECOGNIZED EMOTIONAL OUTPUT

Calm, neutral, surprise, happy, sad, angry, fearful, disgust are primary emotions detected in this speech emotion detection.

IV. METHODOLOGY

SVM: Support Vector Machine (SVM) comes under supervised machine learning category it was used for classification as well as for regression problems also. It is used to map the low dimensional feature vector to high-dimensional feature vector space, so it can solve nonlinear separable problem. SVM has been widely used in pattern classification[12].SVM upholds vector machines. For data comprising of choices set by the login name, A SVM shows the model demonstrating another progressive model. Assume there are just two classifications, which frequently showup in the SVM list.

SVM Linear Classifier:

As far as enrollment, we won't permit an instructor to set a model at home. Zeroing in on this data is planned to beat any issue. Hyperplane plans to be separated into two sections. The most compelling thing to do when arranging a hyper plane is to diminish the confinement from the hyper plane to the closest data in two phases. A hyper-plane demonstrates a huge hyper-plane.

SVM Non-Linear Classifier:

Our information base is generally utilized all over the planet. Admittance to this data in different types of hyperplants ought not be underestimated. Therefore, Vapnik has recommended that you sort the line utilizing a super level stem. In the SVM list without a line, the fundamental data is broken.

We import train_test_split from sklearn model selection. We are using SVC from svm classifier because it is a classification. We set kernel to linear and the train our model.

Description of SVM boundaries:

In this part, we will figure out how to pick the best hyperplan to work with. We will show you Phase 2 data. Classes are introduced in triangles and wheels.

Case 1:

- Analyze the issue in Figure 2 and the data in the two extraordinary segments. At the present time, we want to check out the right airplane that can separate between the two classifications.
- Now, see Figure 1. From the choices to get the right hyperplane. In SVM, we attempt to decide the distance between the hyperplane and the connected data. This is known as the shore.
- The 1 order has a cutoff with the goal that the distance between the left and the right is enormous. So, the finish of our hyperplane will be "first".

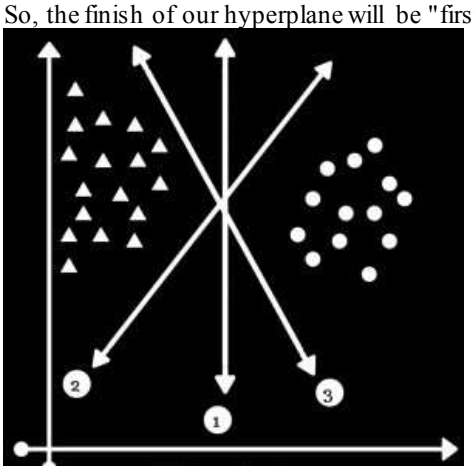

Fig. 1. Support Vector Machine

Case 2:

- In Figure 2, we center around two unique sorts of media. At this moment, we want to find the right hyperplane that can isolate the two classifications.
- Convey data in every class to the left or right. We can choose a hyperplan that can separate between classes for outrageous contrasts.
- Given a decision range as of now, choice 1 shows that there is a distinct contrast between bigtriangleup and \ bigcirc.

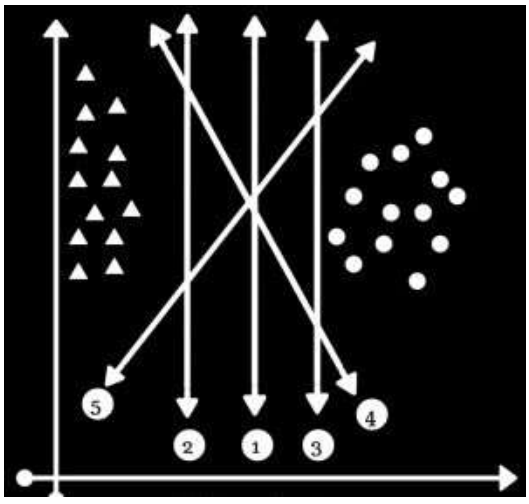

Fig. 2. Support Vector Machine

MLP **:** Multilayer percepton is a part of artificial neural network(ANN). Back propagation is the method used for training a MLP classifier.

Process for building MLP classifier has following steps:

- Declare MLP classifier by defining and the required parameters.
- To train classifier we are giving data to Neural Networks.
- To predict the output we use trained network.

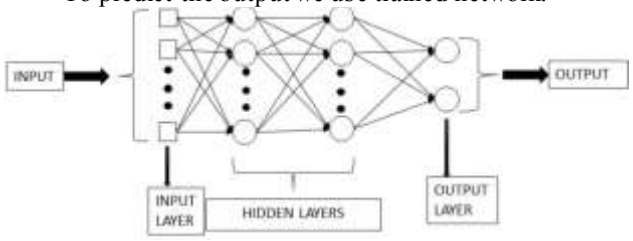

Fig. 3. Multilayered Perceptron

V. SYSTEM ARCHITECTURE

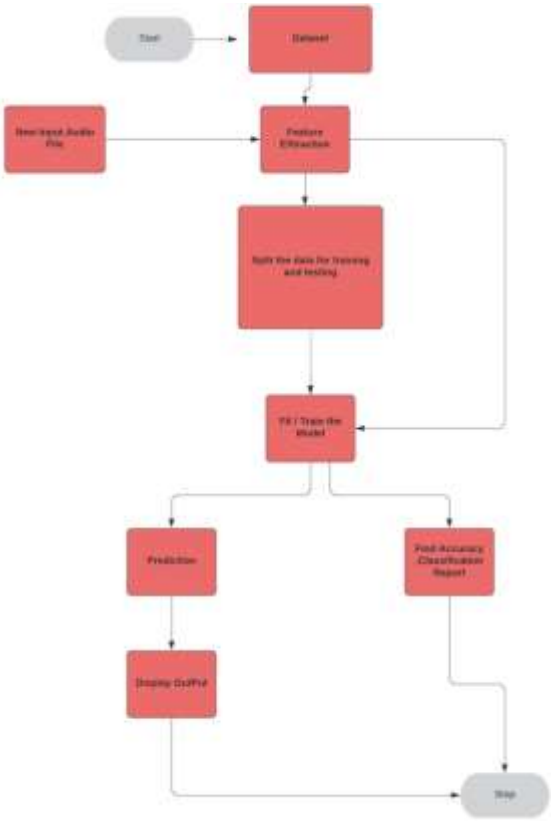

Fig. 4. Training Process Work Flow

VI. DATASETS

Fig. 5. RAVDESS dataset.

This dataset is a combination of song files and speech files. we have a total of 24 actor files in both speech and song[13]. In 40 audio files in song files. The emotions are labelled as follows: 01-'neutral', 02-'calm', 03-'happy', 04-'sad', 05-'angry', 06- 'fearful', 07-'disgust', 08 -'surprised'.

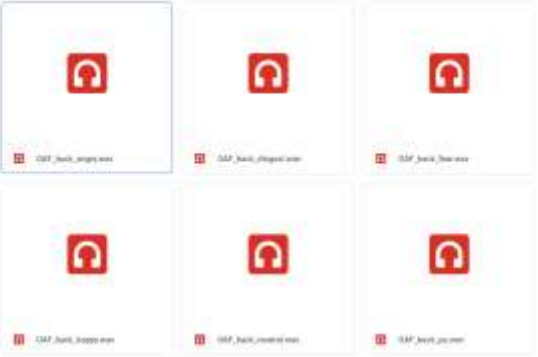

Fig. 6. TESSdataset.

This dataset contains a total of 2800 audio files.

VII. RESULTS

TABLE I. CONFUSION MATRIX

	Predicted_angry	Predicted_sad	Predicted_neutral	Predicted_ps	Predicted_happy
True_angry	92.307693	0.000000	1.282051	2.564103	3.846154
True_sad	12.820514	67.948715	3.846154	6.410257	8.974360
True_neutral	3.846154	8.974360	82.051285	2.564103	2.564103
True_ps	2.564103	0.000000	1.282051	83.333328	12.820514
True_happy	20.512821	2.564103	2.564103	2.564103	71.794876

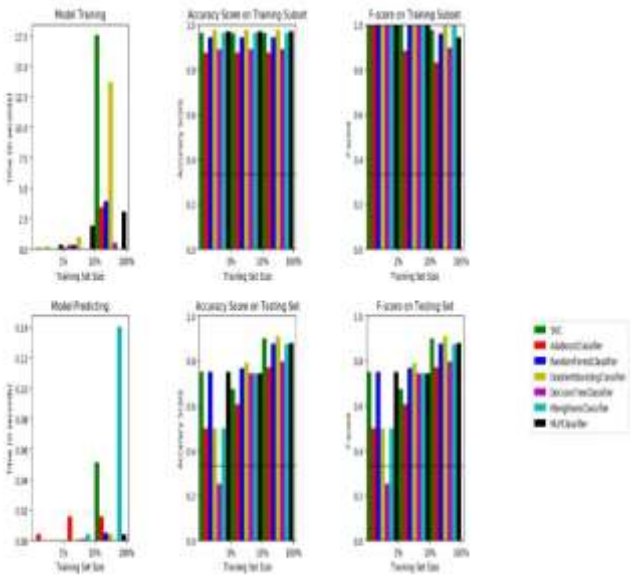

Fig. 7. Histograms on different classifiers.

VIII. CONCLUSION

SER is an interesting subject in software engineering research. The proposed framework resembles the present status of the SER calculation. Later on, the proposed framework could be extended to give multilingual Emotion. Likewise, feelings can be extended to characterize minute levels and design.

REFERENCES

[1] Kottursamy, Kottilingam. "A review on finding efficient approach to detect customer emotion analysis
using deep learning analysis." Journal of Trends in Computer Science and Smart Technology 3, no. 2
(2021): 95-113.

[2] Thakur, Amrita, Pujan Budhathoki, Sarmila Upreti, Shirish Shrestha, and Subarna Shakya. "Real Time
Sign Language Recognition and Speech Generation." Journal of Innovative Image Processing 2, no. 2
(2020): 65-76.

[3] Kaur, Jasmeet, and Anil Kumar. "Speech Emotion Recognition Using CNN, k-NN, MLP and Random
Forest." In Computer Networks and Inventive Communication Technologies, pp. 499-509. Springer,
Singapore, 2021.

[4]Gamage, Kalani Wataraka, Vidhyasaharan Sethu, Phu Ngoc Le, and Eliathamby Ambikairajah. "An i-vector gplda system for speech based emotion recognition." In *2015 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA)*, pp. 289-292. IEEE, 2015.

[5]Han, Jing, Zixing Zhang, Fabien Ringeval, and Björn Schuller. "Reconstruction-error-based learning for continuous emotion recognition in speech." In *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pp. 2367-2371. IEEE, 2017.

[6] Akrami, N., F. Noroozi, and G. Anbarjafari. "Speechbased emotion recognition and next reaction prediction." In *25th Signal Processing and Communications Applications Conference, Antalya*, pp. 1-6. 2017.

[7] Rieger, S. A., Muraleedharan, R., & Ramachandran, R. P. (2014, September). Speech based emotion recognition using spectral feature extraction and an ensemble of kNN classifiers. In *The 9th International Symposium on Chinese Spoken Language Processing* (pp. 589-593). IEEE.

[8] Tabatabaei, Talieh S., and Sridhar Krishnan. "Towards robust speech-based emotion recognition." *2010 IEEE International Conference on Systems, Man and Cybernetics*. IEEE, 2010.

[9] Z. Li, "A study on emotional feature analysis and recognition in speech signal," Journal of China Institute of Communications, vol. 21, no. 10, pp. 18–24, 2000.

[10]T. L. Nwe, S. W. Foo, and L. C. de Silva, "Speech emotion recognition using hidden Markov models," Speech Communication, vol. 41, no. 4, pp. 603–623, 2003.

[11] Vaijayanthi, S., and J. Arunnehru. "Synthesis Approach for Emotion Recognition from Cepstral and
Pitch Coefficients Using Machine Learning." In International Conference on Communication, Computing
and Electronics Systems, p. 515.

[12] Y. Kim, H. Lee, and E. M. Provost, "Deep learning for robust feature generation in audio-visual emotion recognition," in Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP '13), Vancouver, Canada, 2013.

[13].Livingstone, Steven R., and Frank A. Russo. "The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English." PloSone 13, no. 5 (2018): e0196391

[14] Chourasia, Mayank, Shriya Haral, Srushti Bhatkar, and Smita Kulkarni. "Emotion recognition fromspeech signal using deep learning." Intelligent Data Communication Technologies and Internet of Things:Proceedings of ICICI 2020 (2021): 471-481.

[15] R. A. Khalil, E. Jones, M. I. Babar, T. Jan, M. H. Zafar and T. Alhussain, "Speech emotion recognition using deep learning techniques: A review", IEEE Access, vol. 2, no. 7, pp. 117327-117345, 2019.