IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 2, May 2022
Impact Factor: 6.252

Speech Emotion Recognition Using Classifiers and XGBoost Algorithm

Aakanksha Jaykar[1], Akshay Dhaybar[2], Shrinivas Koli[3], Avishkar Shinde[4], Mr. Swapnil N. Pati[5]
Students, Department of Computer Science and Engineering[1,2,3,4]
Professor, Department of Computer Science and Engineering[5]
Prof., Smt. Kashibai Nawale College of Engineering, Pune, Maharashtra, India
aakanksha.jaykar@gmail.com[1], akshaydhaybar2301@gmail.com[2],
shrinivaskoli026@gmail.com[3], friends4ever.avi@gmail.com[4]

Abstract: *Communication is the key to specific one's thoughts and ideas clearly. The time of Machine Learning(ML) is rapidly advancing in bringing more intelligent systems available for everyday use. Intelligent applications are interactive and need minimum user effort to function, and mostly function on voice-based input. A speech percept can reveal information about the speaker including gender, age, language, and emotion. Several existing speech recognition systems employed in ML are integrated with an emotion detection system in order to investigate the spirit of the speaker. The performance of the emotion detection system can greatly influence the performance in many ways and might provide many advantages over the functionalities of those applications. During this proposed project, we perform speech data analysis on speaker discriminated speech signals to detect the emotions of the individual speakers involved within the conversation. We are analyzing different techniques to perform speaker discrimination and speech analysis to seek out efficient algorithms to perform this task.*

Keywords: XGBoost Algorithm

I. INTRODUCTION

For several years now, the expansion within the field of Machine Learning (ML) has been accelerated. ML, which was once a topic understood by computer scientists only, has now reached the house of a typical man within the kind of intelligent systems. The advancements of AI have engendered to many technologies involving Human-Computer Interaction (HCI). going to develop and improve HCI methods is of paramount importance because HCI is that the front-end of ML which ample users experience. a number of the prevailing HCI methods involve communication through touch, movement, hand gestures, voice and facial gestures [1]. Among the various methods, the voice-based intelligent devices are gaining popularity during a wide selection of applications. In an exceedingly voice-based system, a computer agent is required to completely comprehend the human's speech percept so as to accurately acquire the commands given to that.

This field of study is termed as Speech Processing and consists of three components: identification Speech Recognition Speech Emotion Detection Speech Emotion Detection is challenging to implement among the opposite components thanks to its complexity. Furthermore, the definition of an intelligent system requires the system to mimic human behaviour.

II. MOTIVATION

Identifying the emotion expressed in an exceedingly speech percept has numerous use instances withinside this day packages. Human-Computer Interaction (HCI) could be a discipline of studies that research interactive packages among people at large and computers. For a powerful HCI application, it's miles necessary for the pc gadget to apprehend extra than simply words.

On the other hand, the sphere of Internet of Things (IoT) is swiftly growing. Many actual phrase IoT packages that are used on a each day foundation inclusive of Amazon Alexa, Google Home and Mycroft characteristic on voice-primarily based totally inputs. The position of voice in IoT packages is pivotal. The examine in a very current article foresees that through 2022, approximately based on voice instructions only. These voice interactions is mono-directional or bi-directional, and in each instances, it's miles rather essential to recognize the speech signal. Considering emergency conditions wherein the buyer is ineffective to absolutely offer a voice command, the emotion expressed thru the consumer's tone of voice could

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 2, May 2022
Impact Factor: 6.252

also be wont to show on certain emergency capabilities of the vehicle. A plenty easier application of speech emotion detection could also be visible in name only centers, in which computerized voice calls is also correctly transferred to customer support retailers for similarly discussion.

III. Proposed Methodology

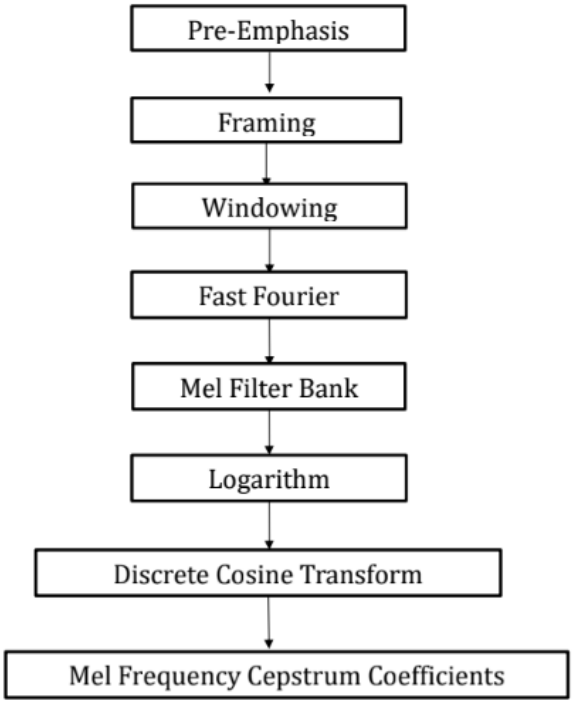


Figure 1: Stages involved in MFCC Feature Extraction

Here we take up the task in two different sections: speech and speaker recognition and emotion recognition. within the case of speech and speaker recognition the concept of MFCC alone suffices and provides us a satisfactory output. it's worth noting that out database within the case of speech and speaker recognition is computerized voice, and hence chances of noise and interference is incredibly low. But using real time data, as shown in [3] illustrates how noise can distort MFCC outputs. Thus requirement of compensatory methods won't be crucial, but it is recommended within the case of emotion recognition, features like pitch, amplitude and also the coefficients are taken, and therefore the means and variances of the sets of knowledge are grouped, thereby minimising the clutter without affecting accuracy.

2.1 Pre Emphasis

The speech signal represented by x [n] is then sent to a highpass filter as given within the below equation.

$$y[n] = x[n] - ax[n-1]$$

where y [n] is that the output. the price of a which is normally from 0.9 to 1.0. The Z transform of this equation is given by:

$$H(z) = 1\text{-}az\text{-}1.$$

This progression makes up for the high frequency a component of the speech signal which beforehand got suppressed due to the human component of sound generation. It can likewise enhance the importance of high-frequency formants.

2.2 Framing

This process involves segmenting the speech test into a tiny low frame acquired from analog to digital conversion. The speech signal is split into frames of M samples, with adjacent frames being separated by N where N could be a smaller amount than M. the primary frame consists of the first M tests. The second frame has N samples after the first frame which is overlapped by M – N samples. Each frame overlaps with two other resulting frames.

The technique is thus called framing.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 2, May 2022
Impact Factor: 6.252

2.3 Fast Fourier Transform

The progression is employed to induce the magnitude frequency response of each frame. it's viewed as that the signal within a frame is periodic and continuous when FFT is performed. Regardless of the likelihood that the signals are discontinuous then also it'll be performed but the frame's first and last discontinuous points will produce unwanted effects within the frequency response. to resolve this issue, each frame is duplicated by a hamming window to increase its continuity

2.4 Mel Filter Bank Processing

The speech signal doesn't follow the linear scale and also the frequency range is extended. The yield is given by the complete of its filtered spectral components.

2.5 Discrete cosine Transform (DCT)

This process is employed to convert from frequency domain into time domain using Discrete Cosine Transform (DCT). The result provides Mel Frequency Cepstrum Coefficients (MFCC).[5] The representation of local spectral properties of the signal along with frame analysis is good when done by cepstral analysis.

The cepstral coefficient is also filtered by the auditory filters like bark and Mel filters. The cepstral features are derived from the filter bank designed as per the sensory system. It depends on pitch observation and also the filters are placed in triangular shaped.

2.6 Emotion recognition using SVM

Emotion recognition requires a training period where features of a database are collected and congregated to extract viable information. Once a database is ready features are extracted from the test signal, comparing with formed hyperplanes and allocating it to the foremost matched group.

SVM [7] is one in every of the classifier utilized for the emotion recognition. It recognizes the pattern and analyzes the knowledge. it's one amongst the foremost popular classifiers used for emotion database classification.

III. TRAINING TESTING

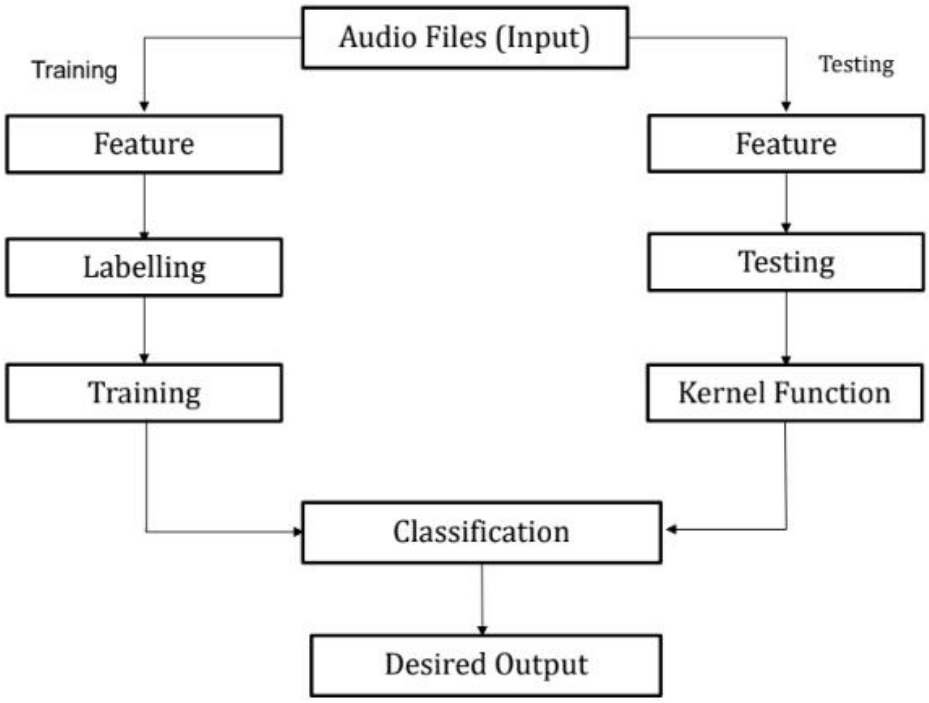


Figure 2: Block Diagram of Emotion Recognition

Figure 2 is an example of classification into two groups.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 2, May 2022
Impact Factor: 6.252

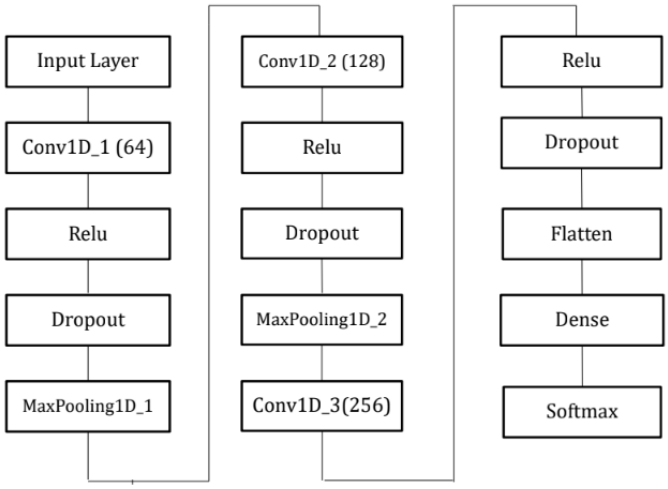
3.1 CNN

The deep neural community(CNN) designed for the category venture is pronounced operationally in Fig. 1. The community can paintings on vectors of forty capabilities for every audio record supplied as enter. The forty values constitute the compact numerical shape of the audio body of 2s length. Consequently, we offer as enter a of length < number> x forty x 1 on which we finished one spherical of a 1D CNN with a ReLu activation characteristic, dropout of 20%, and a max-pooling characteristic 2 x 2.

3.2 CNN Model

The rectified linear unit (ReLu) may be formalized as g(z) = max, and it permits us to attain a huge fee in case of activation with the aid of using making use of this characteristic as a terrific desire to symbolize hidden units. Pooling can, on this case, assist the version to cognizance simplest on most important traits of each part of data, making them invariant with the aid of using their position. We have run the method defined another time with the aid of using converting the kernel length.

Following, we've got implemented some other dropout after which flatten the output to make it well suited with the subsequent layers. Finally, we implemented one Dense layer (absolutely related layer) with a softmax activation characteristic, various the output length from 640 factors to eight and estimating the opportunity distribution of every of the lessons well encoded (0=Neutral; 1= Clam; 2= Happy; Sad=3; Angry=4; Fearful= 5; Disgust=6; Surprised=7).



3.3 XGBoost

The XGBoost is having a tree getting to know set of rules in addition to linear version getting to know, and due to that, it could do parallel computation on a unmarried system. When we examine XGBoost with different gradient boosting algorithms, XGBoost seems to be clearly fast, about 10 instances quicker than different implementations.

The XGBoost set of rules makes use of the gradient boosting selection tree set of rules.The gradient boosting approach creates new fashions that do the venture of predicting the mistakes and the residuals of all of the earlier fashions, which then, in turn, are delivered collectively after which the very last prediction is made.

3.4 MLP

A multilayer perceptron (MLP) is a category of feed forward synthetic neural network (ANN).MLP makes use of a supervised gaining knowledge of method known as back propagation for training. Its a couple of layers and non-linear activation distinguish MLP from a linear perceptron. It can distinguish information that isn't always linearly separable. A multilayer perceptron (MLP) is a category of feedforward synthetic neural network (ANN).

MLP makes use of a supervised gaining knowledge of method known as backpropagation for training. Its a couple of layers and non-linear activation distinguish MLP from a linear perceptron. It can distinguish information that isn't always linearly separable.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 2, May 2022
Impact Factor: 6.252

3.5 SVM

"Support Vector Machine" (SVM) is a supervised gadget mastering set of rules that may be used for each class or regression challenges. However, it's miles often utilized in class problems.

In the SVM set of rules, we plot every statistics object as a factor in n-dimensional space (in which n is the quantity of capabilities you have) with the price of every function being the price of a specific coordinate. Data may be scaled earlier than making use of to an SVM classifier to keep away from attributes in extra numeric stages at the same time as processing it. Scaling additionally serves the motive of keeping off a few numerical problems for the duration of the calculation.

IV. CONCLUSION

We study the use of Mel Frequency Cepstrum Coefficients in speech and speaker recognition. We also bear variations applicable, like extracting coefficients after performing an FFT, or utilizing different windows so as to raised the functioning and minimize the error. We follow up with automatic emotion recognition; discuss the approaches available to serve the aim, and zeroing in on one method, we discuss characteristics of Support Vector Machines within the segregation and classification of various aspects of speech that are extracted, like amplitude, pitch etc. that are essential to know the speaker's state. With this we are able to acquire a completely functional recognition system that may be used for security based systems or the other variety of human-machine interactions (HMI).Our evaluation shows that the proposed approach yields accuracies of 86%, 84% and 82% using CNN, MLP classifier and SVM classifiers, respectively, for 8 emotions using Ryerson Audio-Visual Database of Emotional Speech and Song(RAVDESS) dataset and Toronto Emotional Speech Set (TESS) Dataset.

FUTURE WORK

For future advancements, the proposed project are often further modeled in terms of efficiency, accuracy, and usefulness. A sentiment detection using vocabulary, may be integrated with speech emotion detection to spot a possible sarcasm. Therefore, within the future, there would emerge many applications of a speech-based emotion recognition system.

REFERENCES

[1]. Koren, Leon, and Tomislav Stipancic. "Multimodal Emotion Analysis supported Acoustic and Linguistic Features of the Voice." In International Conference on Human-Computer Interaction, pp. 301-311. Springer, Cham, 2021.

[2]. Prasomphan, Sathit. "Detecting human emotion via speech recognition by using speech spectrogram." In 2015 IEEE International Conference on Data Science and Advanced Analytics (DSAA), pp. 1-10. IEEE, 2020.

[3]. Ibrahim, Hemin, Chu Kiong Loo, and Fady Alnajjar. "Speech Emotion Recognition by Late Fusion for Bidirectional Reservoir Computing with Random Projection." IEEE Access (2021).

[4]. Brownlee, J. (2018). a mild Introduction to the Gradient Boosting Algorithm for Machine Learning - Machine Learning Mastery. [online] Machine Learning Mastery. Available at: https://machinelearningmastery.com/gentle-introduction-gradient-boosting algorithm- machine learning/.

[5]. Ray, S. (2018). Decision Tree — Predictive Analytics. [online] Analytics Vidhya. Available at: https://www.analyticsvidhya.com/blog/2015/01/decision- treesimplified/2/.

[6]. Srivastava, T. (2018). Introduction to KNN, K-Nearest Neighbors : Simplified. [online] Analytics Vidhya. Available at: https://www.analyticsvidhya.com/blog/2014/10/introduction-kneighboursalgorithm- clustering/

[7]. DeZyre. (2018). Principal Component Analysis Tutorial. [online] Available at: https://www.dezyre.com/data-science-in-pythontutorial/principal-component- analysistutorial.

[8]. En.wikipedia.org. (2018). Bayes' theorem. [online] Available at: https://en.wikipedia.org/wiki/Bayes En.wikipedia.org. (2018). Logistic regression. [online] Available at: https://en.wikipedia.org/wiki/ Logisticregression.