# SPEECH EMOTION RECOGNITION IMPLEMENTATION

DHARANI PRASAD S[1], BRIJ VISHAL RAJPUT[2], DHANJIT DEKA[3], ADITHYA BALAGOPAL[4], DR POOJA NAYAK S[5]

[123456] prasadsdharani@gmail.com, brijvishal.bv@gmail.com, dhanjit1951@gmail.com, adithyab0727@gmail.com, pooja-ise@dsatm.edu.in

[12345] Student, Department of Information science and engineering, DSATM, Bangalore-88, Karnataka

[6]Faculty, Department of Information science and engineering, DSATM, Bangalore-88, Karnataka

*ABSTRACT*

Dialogue is one of the most common ways that people communicate. We rely on it heavily to the point that we understand its importance while using other forms of correspondence, such as SMS and instant conversations, where we regularly utilize emoticons to convey the feelings associated to the message. Emotions are crucial to the mental health of humans. It serves as a vehicle for communicating one's viewpoint or mental condition to others. Speech Emotion Recognition (SER) is the process of deriving the speaker's emotional state from the speech signal. Any intelligent system with limited processing resources may be trained to recognize or synthesize the few universal emotions—Neutral, Anger, Happiness, and Sadness—as needed. Because both spectral and prosodic traits include the emotional information, they are both utilized in this study to identify speech emotions. One of the spectral properties is the MFCC, or mel-frequency cepstral coefficients. Prosodic variables such as fundamental frequency, loudness, pitch, intensity of speech, and glottal factors are utilized to describe various emotions. Each utterance's potential characteristics are taken out in order to create a computational mapping between speech patterns and emotions.

Keywords: speech motion recognition; emotion recognition; automatic speech recognition; machine learning, MFCC;

## 1. INTRODUCTION

In recent times, the significance of understanding human speech emotions has grown in order to enhance the effectiveness and naturalness of human-machine interactions. The difficulty in differentiating performed and natural emotions makes it a very difficult task to recognise human emotions. Speech Emotion Recognition, often known as SER, is an example of an attempt to recognise human emotion and complex emotional states from speech. This takes use of the way tone and pitch of speech typically convey concealed emotion. This is also the miracle that allows animals like dogs and horses to have the capacity for comprehension human emotion. Since sentiments are ephemeral and expressing sound is challenging, SER is intense. In this project, we'll build a model using an MLP Classifier using the libraries librosa, soundfile, and sklearn, among others. This genuinely want to experience emotions through sound recordings. The data will be stacked, the highlights will be extracted, and then the dataset will be divided into preparation and testing sets. The model will then be trained after which an MLP Classifier will be implemented. Finally, we'll check the accuracy of our model. This genuinely want to experience emotions through sound recordings. After stacking the data and eliminating any highlights, the dataset will be divided into preparation and test sets. Applications for voice emotion detection may be found in a variety of settings, including contact centres and human-computer interaction.

Humans are capable of automatically and unconsciously recognising emotions. Since it is a crucial step in human-to-human communication, emotions must be taken into account for improved human-machine connection.

Emotions can be quantified using one of three main methods: categorical, continuous, or appraisal-based. Speech is a complicated signal made up of a variety of pieces of data, including details about the user, the message being sent, the language, the context, the emotions, etc. One of the key areas of digital signal processing is speech processing, which has applications in security, telecommunication, assistive technology, human computer interfaces, and other areas. In order to establish a natural relationship between a machine and a human, speech emotion recognition is crucial. Speech emotion recognition is extracting a speaker's emotional state from their speech. The voice signal's acoustic aspect is called Feature. A small portion of information from the speech signal is extracted using the feature extraction method so that it may subsequently be utilised to identify each speaker.

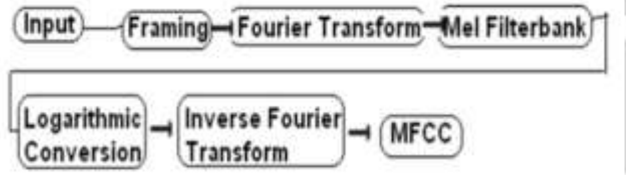
## 2. TECHNOLOGY

### 2.1 MACHINE LEARNING

A developing technique called machine learning makes it possible for computers to learn autonomously from historical data. Machine learning employs a variety of techniques to create mathematical models and make predictions based on previous knowledge or data. Currently, it is utilized for
many different things, including recommender systems, email filtering, Facebook auto-tagging, picture identification, and speech recognition. Four main advancements make up our SER framework. The voice test selection comes first. the second highlights vector that the highlights have separated into. As the next step, we tried to determine which details are often relevant to differentiate each emotion. These features
are known to the AI classifier for recognition.

### 2.2 MFCC

A representation of a sound's short-term power spectrum used in sound processing is called a mel-frequency cepstrum (MFC), which is based on a linear cosine transform of a log power spectrum on a nonlinear mel scale of frequency. An MFC is made up of a number of coefficients known as mel-frequency cepstral coefficients (MFCCs).They are produced from a nonlinear "spectrum-of-a-spectrum"cepstral representation of the audio sample. The mel-frequency cepstrum (MFC) differs from the cepstrum in that the frequency bands are evenly spaced on the mel scale, which more closely resembles the response of the human auditory system than the linearly-spaced frequency bands used in the conventional spectrum.



### 2.3 MULTI LAYER PERCEPTRON

A kind of feedforward artificial neural network (ANN) is called a multi-facet perceptron (MLP). The term "MLP" is loosely used to refer to any feedforward ANN, but it is also also used specifically to refer to networks made up of different layers of perceptron’s (with edge initiation). Occasionally, multi-facet perceptrons are referred to as "vanilla" neural networks, especially when they include a single secret layer. At least three levels of hubs make up an MLP: an info layer, a hidden layer, and a yield layer. Every hub—aside from the information hubs—is a neuron that makes use of a nonlinear initiation work. Backpropagation is a controlled learning technique that MLP employs while planning. MLP may be distinguished from a straight perceptron by its multiple layers and non-direct actuation. It is capable of recognising data that is not immediately separable. MLPs are useful in research because they can handle problems stochastically, which frequently enable approximated responses for highly complicated problems like health guess. According to Cybenko's premise, MLPs are widely used capacity approximators, and relapse analysis can use them to create numerical models. MLPs are excellent for classifier computations since order is a particular instance of relapse when the response variable is absolute.

### 2.4 RAVDESS

7356 recordings totaling 24.8 GB make up the Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS). A neutral North American intonation is used by the 24 experienced entertainers (twelve male and twelve female) who express two lexically coordinated with articulations in the data store. Discourse includes expressions that are calm, happy, upbeat, angry, sad, shocked, and the song evokes sentiments of serenity, joy, pity, fury, and misfortune.

### 2.4 MEL SPECTROGRAM

The spectrogram is created by doing a Fast Fourier Transform on overlapping windowed portions of the data. This spectrogram merely shows amplitude that has been mapped on a Mel scale.

### 2.5 CHROMA

In a standard chromatic scale, a chroma vector is a 12-element feature vector that represents the energy of each pitch class in the signal.

## 3. TOOLS AND PLATFORMS USED A. JupyterLab

The electronic user interface for Project Jupyter, known as JupyterLab, is open-source and incorporates all of the key features of the Jupyter Notebook, including a scratch pad, terminals, content tools, recording programmes, rich outputs, and more. However, it also provides better support for outside extensions.

### B. Pip install

A Python-based framework called Pip is used to introduce and manage programming packages. It connects to the Python Package Index, an online repository for free and paid private bundles.

### C. Librosa

A Python library for studying sound and music is called Librosa.It has a complement bundle design, standardises names and interfaces, reverse similarity, isolated capacity, and understandable code.

### D. Sklearn

The most useful Python AI library is probably scikit-learn. Characterization, relapse, bunching, and dimensionality reduction are just a few of the very effective tools for AI and factual showing that are available in the Sklearn toolkit.

### E. PyAudio

The cross-stage sound I/O library, PortAudio, is connected to Python using PyAudio. We can surely play and record sound using Python and PyAudio on a variety of stages.

### F. NumPy

The Python package NumPy is used while working with exhibits. Additionally, it has the ability to operate in the fourier, direct variable, and framework spaces.

### G. Sound File

A sound library called SoundFile is reliant on CFFI, NumPy, and libsndfile. Libsndfile, a free, cross-platform, open-source (LGPL) library for reading and creating numerous analysed sound document formats that experiences abrupt increases in demand for several stages including Windows, OS X, and UNIX, supports record reading and composition.

## 4. PROPOSED METHODOLOGY

Some time recently beginning the method, the sound tests are passed to a sexual orientation reference database utilized for sexual orientation location. Factual strategies utilize pitch as a sexual orientation acknowledgment include .

A reference database can be utilized to decide the upper and lower pitches for both male and female tests . The primary step was to part the input human voice test into outlines with a outline estimate of 16 ms. In a later step, this was done for frame-level classification. The most include for feeling segregation in each outline was the MFCC (Mel-Frequency Cepstral Coefficient). A reference database is kept up at the MFCC for numerous feelings such as pity, outrage, lack of bias, and joy. I compared the MFCCs of the outlines.

### 4.1 Basic Module Descriptions

Our SER system comprises of four crucial advancements. Dialect test determination comes to begin with. The moment highlight vector with the highlights expelled is boxed. As a following step, we attempted to discover out which points of interest are frequently pertinent to recognize person feelings. AI classifier recognizes and recognizes these highlights.

### A. Feature Extraction

There are numerous boundaries to talk signals that reflect an eager quality. The thought that we ought to utilize accentuation is one of the persevering subjects in recognition. Various common highlights such as vitality, pitch, formants, and a few extend highlights such as Straight Desire
Coefficients (LPC), Mel-Frequency Cepstrum Coefficients (MFCC), and direction ghost highlights are accessible in continuous investigate. confined. In arrange to dispense with the hot highlights in this work, the ghostly highlights are balanced with MFCC.

### B. Feature Choice

The objective is to make strides gathering exactness on a specific errand for a specific learning computation. The security affect is to diminish the number of highlights that trigger the final arrange demonstrate. Highlight determination (FS) points to choose a subset of the foremost critical highlights from the starting highlights based on a certain pertinence score.

This as a rule moves forward acknowledgment exactness. Without a doubt, it can decrease the execution time of the preparing computation.

### C. Classification Methods

Discrete gathering of sensations was accomplished utilizing various AI computations. The reason of these calculations is to memorize from hone tests and apply that information to organize unused points of view. There's no authoritative reply to the learning computation choice. Each strategy has its possess qualities and shortcomings. Hence, in this case, we choose to consider seeing three distinctive classifiers. The leading edge classifiers in AI are called bolster vector machines (SVMs) . It has moreover been broadly utilized in a few considers on the detection of sonic feelings. Compared to other classifiers, it bunches exceptionally well, particularly when there's small preliminary information.

4.2 Schema Design - Data Integrity and Constraints/Dataset Creation Mel-Frequency Cepstrum Coefficients (MFCC) are the foremost common way to depict the ghostly properties of chimes. For talk acknowledgment, it is suitable to consider how open human discernment is to these frequencies. At the scale after cycles, the Lobby transformer and vitality extend of each edge were examined and plotted. The primary 12 DCT coefficients of the Mail Log Vitality Partitioned Cosine Transform (DCT) assessment given the MFCC values used within the approval intelligent. In our consider, we confined the primary 12 prerequisites for the MFCC coefficients when looking at discourse signals at 16 KHz. Calculate the cruel, contrast, standard deviation, kurtosis, and skewness of each prerequisite coefficient, and this applies to all other coefficients.

4.3 MFCC approach

The input for the proposed work is the human voice. The input is at that point part into outlines with each 50ms coming about in 10ms of information cover. This can be since no information is lost. The pitch autocorrelation work is utilized to calculate the basic recurrence.

Based on which sexual orientations can be classified, an normal affinity esteem is calculated utilizing the Bolster Vector Machine reference database.

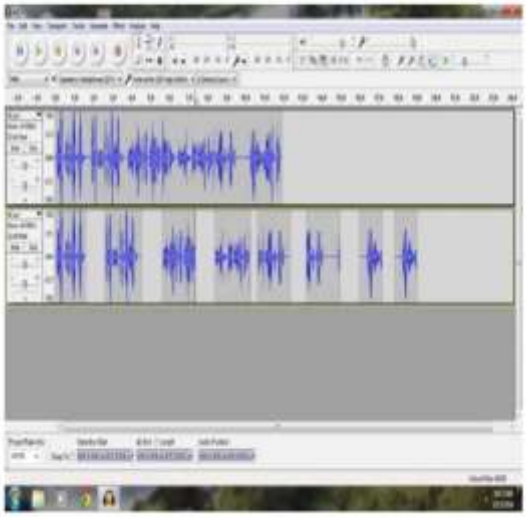


Figure 2. Frames of Data

Each outline can be input into the proposed MFCC strategy for feeling acknowledgment. The human discourse mel-frequency cepstrum coefficient work comprises of a set of four steps. Each is subjected to a Quick Fourier Change to decide the least and greatest frequencies. After utilizing covering triangular windows to outline the control of the range gotten over, the mel channel bank can be utilized to decide the size values. Her MFCC values for each outline are decided after applying a discrete cosine change to recover information misplaced amid sound clip compression .

We are able at that point obtain the normal opinion score by preparing these scores employing a backpropagation organize and a outspread premise work arrange.

We utilized a Gaussian enactment work within the covered up layer and an character enactment work within the output layer, and set the learning rate of the RBF network to 0.001 for organize preparing.

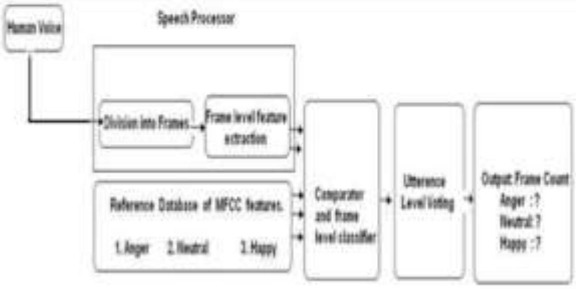

Human Voice
Speech Processor
Division into Frames
Frame level feature extraction
Reference Database of MFCC features.
1. Anger 2. Neutral 3. Happy
Comparator and frame level classifier
Utterence Level Voting
Output: Frame Count
Anger : ?
Neutral: ?
Happy : ?

Figure 3 . Standard Approach for Emotion Classification

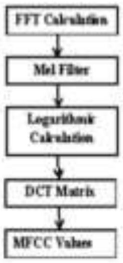


Figure 4 . Proposed MFCC Approach

A slope enactment work within the covered up layer and a twofold step work within the yield layer set the learning rate of the BPN organize to 0.001 for arrange preparing. At long last, these values can be compared to the Phonetic Estimation Reference Berlin Database to classify the estimation set.

## 5. CONCLUSION

Numerous research and surveys have examined the use of deep learning, machine learning, and picture categorization approaches for emotion recognition. Understanding emotions is an indispensable component of transdisciplinary applications. Given robots are incapable of comprehending a speaker's sentiments on their own, emotion recognition of speech has long been an essential niche of research in systems incorporating human-machine interaction.The notion of emotion recognition using the MFCC method and Radial Basis Function Network has been put into practise in this paper. The MFCC approach for emotion recognition from speech is a standalone method that yields more accurate results without calculating any supplementary acoustic parameters. Several databases have been investigated for this project's attempt to employ inception net to solve the problem of emotion recognition.

the KAGGEL and RAVDESS databases were used as the dataset for this experiment. Utilised TensorFlow to train this model. Accuracy rate of about 85% is achieved.

## 6. FUTURE WORK

The application of deep learning algorithms for emotion recognition has been the subject of numerous studies and surveys. In the future, it will be vital to have a system like this that is considerably more reliable and has countless applications across all industries. We will work to increase our accuracy score to over 90% in the future. In the future, we'll also strive to incorporate real-time recording and real-time prediction. We will attempt to use an ensemble of lexical and acoustic models in conjunction with a lexical features-based approach to SER. This will increase the system's accuracy because emotional expression might sometimes be nonverbal or contextual. In the future, we'll also make an effort to differentiate between male and female voices.

## 7. REFERENCES

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

[2]Navya Damodar, Vani H Y, Anusuya M A. Voice Emotion Recognition using CNN and Decision Tree. International Journal of Innovative Technology and Exploring Engineering (IJITEE), October 2019.

[3] RAVDESS Dataset: https://zenodo.org/record/1188976#.X5r200g_zZPZ.

[4]MLP/CNN/RNN Classification: https://machinelearningmastery.com/when-to-use-mlp-cnn-and-rnn-neural-networks/.

[5] MFCC: https://medium.com/prathena/the-dummys-guide-to-mfcc-aceab2450fd.

[6]M. Neumann and N.T. Vu, "Improving Speech Emotion Recognition with Unsupervised Representation Learning on Unlabeled Speech," ICASSP 2019 -2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, United Kingdom, 2019.

[7] H. Zheng and Y. Yang, "An Improved Speech Emotion Recognition Algorithm Based on Deep Belief Network," 2019 IEEE International Conference on Power, Intelligent Computing and Systems (ICPICS), Shenyang, China, 2019.

[8] J. Umamaheswari and A. Akila, "An Enhanced Human Speech Emotion Recognition Using Hybrid of PRNN and KNN," 2019 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COMITCon), Faridabad, India, 2019.

[9] Ragot M, Martin N, Em S, Pallamin N, Diverrez JM. “Emotion recognition using physiological signals”: Laboratory vs. wearable sensors. In: International Conference on Applied Human Factors and Ergonomics. Springer,2017.