iJRASET
International Journal For Research in Applied Science and Engineering Technology



INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 9 **Issue:** VIII **Month of publication:** August 2021

DOI: https://doi.org/10.22214/ijraset.2021.37375

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 9 Issue VIII Aug 2021- Available at www.ijraset.com

Speech Emotion Recognition

Prof. Swethashree A[1], Ganesh. T[2], J Aravind[3], M. Venkatratna[4], R. Gayathri[5]
[1, 2, 3, 4, 5]*Department of Computer Science and Engineering, Ballari Institute of Technology and Management, Ballari*

Abstract: ***Speech Emotion Recognition, abbreviated as SER, the act of trying to identify a person's feelings and relationships. Affected situations from speech. This is because the truth often reflects the basic feelings of tone and tone of voice. Emotional awareness is a fast-growing field of research in recent years. Unlike humans, machines do not have the power to comprehend and express emotions. But human communication with the computer can be improved by using automatic sensory recognition, accordingly reducing the need for human intervention. In this project, basic emotions such as peace, happiness, fear, disgust, etc. are analyzed signs of emotional expression. We use machine learning techniques such as Multilayer perceptron Classifier (MLP Classifier) which is used to separate information provided by groups to be divided equally. Coefficients of Mel-frequency cepstrum (MFCC), chroma and mel features are extracted from speech signals and used to train MLP differentiation. By accomplishing this purpose, we use python libraries such as Librosa, sklearn, pyaudio, numpy and audio file to analyze speech patterns and see the feeling.***

Keywords: Speech emotion recognition, mel cepstral coefficient, neural artificial network, multilayer perceptrons, mlp classifier, python.

I. INTRODUCTION

In human-computer communication (HCI), Speech Emotional Recognition (SER) is growing it is important in various programs. Right now, the feeling of speech recognition is an emerging field of practical crossing artificial intelligence and mental functioning; otherwise, it is popular a research topic for signal processing and pattern recognition. Research is widely used in human and computer communication, interactive teaching, entertainment, safety fields, and so on. The process of processing emotions and the expression of awareness it is usually made up of three parts, the first being the speech signal detection, followed by a feature release followed by seeing emotions.

The most prepared method for Speech recognition is a method based on the neural network. Artificial Neural Networks, (ANN) are biologically inspired data processing tools. Speech recognition modeling with artificial neural networks (ANN) do not require precursors knowledge of the process of speaking this method quickly became an attractive place for HMM. RNN can read sub-speech relationships - data and know-how modeling time-dependent phonemes.

Normal neural Multi-Layer Perceptron (MLP) networks have always existed it continues to be used for speech recognition and diversity other speech processing applications. Speech recognition is acoustic signal conversion process, installed by microphone or phone, in set of characters.

They can and serves as an input into the furtherance of language practice in it acquires understanding of speech, a topic covered in a paragraph. As we know, speech recognition performs similar functions the human brain.

II. PRESENT WORK

The removal of the traditional emotional element was based on analysis and comparison of all types of emotional states parameters, you select all the emotional aspects with high sensitivity adjustment for the purpose of removing the feature. The traditional method focuses on the analysis of features in speech such as time construction, size construction, and frequency construction, etc. Time to speak construction refers to the utterance of emotional expression difference in time. Different emotions have different types of call time intervals that can also be detected analyzed by scanning a few data sets. Such a difference can also be found in the frequency and width of parallel audio signal parameters.

This way, however the basic concept of separating emotions from speech, it and there are many problems as the time taken is high, judgment methods can vary, and complex systems are required. There are also many previously proposed models improve the accuracy of SERS predictions. For example, we have Support Vector Machine (SVM), which is a classifier that mathematically includes sound parameters signal so you can predict emotions. This model has been has been very successful in the SER domain. But the main one. The bad thing about SVM's is that it can only split data in two stages; it could be paragraph 1 or 2. And some injustice enter processing time, noise leading to errors in prediction and low accuracy.

©IJRASET: All Rights are Reserved

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 9 Issue VIII Aug 2021- Available at www.ijraset.com

III.PROPOSED WORK

A. Neural Networks

Neural networks are a set of algorithms, which are freely modeled in the background the human brain, designed to recognize patterns. The patterns they see are numerical, contained in vectors, access to all real data, be it images, sound, text or time series, should be translated. It helps us to get together and separate. You can think of it as a layer of integration and separation more than the raw data you store and carry. They help to unlabeled data according to the similarity between sample input, and they split the data when they received the a database with a training label on it. Neural networks appear as an attractive way to model acoustic in ASR later 1980s. Since then, neural networks have been used for many things speech recognition features such as phonetic editing, single word recognition, audio speech recognition, Speaker recognition and speaker view. Neural networks make a few clear assumptions about includes mathematical features than HMMs and has several qualities that make visual models attractive in speech recognition.

1) *Deep Feedforward and Recurrent Neural Networks:* The deep neural feed supply network is the artificial neural a network with multiple hidden layers of units during installation and output layers. DNNs can be complex non-linear models relationships. Its structures form design models, where additional layers enable the formation of features from the bottom layers, which provide greater learning ability and thus be more dynamic modeling complex speech data patterns. One the basic premise of deep learning is to finish hand-crafted engineering and the use of immature features. This the policy was first successfully tested in the construction of deep auto-encoder in "green" spectrogram or filter bank features, indicating its superiority over Mel-Cepstral features that contain a few stages of consistent evolution from spectrograms. True "green" elements of speech, waveforms recently shown to be productive excellent results for speech recognition.
2) *Mel-Frequency Cepstral Coefficients (MFCC):* Mel-frequency cepstral coefficients (MFCC) is one of the most popular sound feature. It is a presentation signals where a feature called a window cepstrum. The short-term signal is found in the FFT of that signal. The signal then goes to the frequency axis of the mel frequency scale using log based transform, and then corresponds to using the converted Discrete Cosine Transform. The measures to remove MFCC features include prior emphasis, frame block and window installation, FFT size, Mel filterbank, log energy, and DCT. MFCC uses melting rate, prepared for the normal response of the human ear. Due to this, the MFCC has been shown to be very important in speech recognition field and tried to associate with it emotional recognition. Depending on the Spectral sound features such as the MFCC is better suited for the division of the N route.
3) *Multilayer Perceptrons Classifier (MLP Classifier):* Subsequent work with multilayer perceptrons has shown that they are able to guess the XOR operators as well many other non-direct functions. Multilayer perceptrons it is often used in supervised learning problems. They train in a collection of input-output pair and learn to model relationships (or dependencies) between those inputs and outcomes. Network so there is a simple translation as an input-output method model, weighted and restricted (discriminatory) free model parameters. Significant issues in MLP formulation enter the specified number of hidden layers and the file the number of units in these layers. Number of units hidden in order usage is unclear. As a good start to using any one layer is hidden, with the number of units equal to half total number of input and output units.

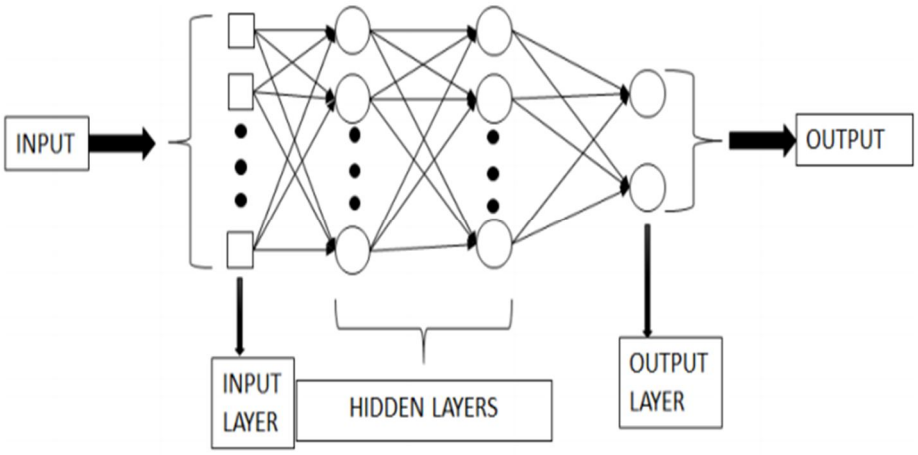


Fig. 1 Multilayered Perceptron

©IJRASET: All Rights are Reserved

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 9 Issue VIII Aug 2021- Available at www.ijraset.com

IV. SPEECH EMOTION RECOGNITION USING MLP CLASSIFIER

In Speech Emotion Recognition System (SER), audio files are provided as input. Data sets go with the file the number of process blocks that make it work help with speech barrier analysis. Details used to switch to appropriate format and various features from audio files are extracted using various steps such as fencing, hamming, windows, etc. The process helps to split audio files into numerical values represent frequency, time, amplitude or other such parameters that may assist in the file analysis of audio files. After the required release features from audio files, model is trained. We have use the RAVDESS dataset for audio files with speech of 24 people for a variety of parameters. Training, we store numerical values of emotions and their variability puts in the same with a different layout. This arrangement exists provided as an addition to the existing MLP Classifier started. The separator identifies the different categories in the file sets the data and separates them with different emotions. The model will now be able to understand the price range of speech parameters fall into certain senses. Because we test the performance of the model, if we include the unknown test dataset as input, it will download the parameters again predict emotion as individual training dataset values. The System accuracy is displayed in the form of percentages which is the end result of our project.

V. IMPLEMENTATION

MLP isolates are used to predict emotions from the feed input. We get results using five published features. We send five different features to the model. Using features independently and completely passing it off we get the best deviations of predictive emotions, such as single input the parameter is not enough to come up successfully prediction. The Ravdess database is transferred to MLP Separator training model, divides the database into 75:25 rate, e.g. training and assessment database. The database contains audio samples of 24 professional North American actors American highlight. Eight types of emotions are covered. The Classifier is used as it works well in time series data, in our case the sound we will be predicting that the emotion. Figure 2 shows the training process.

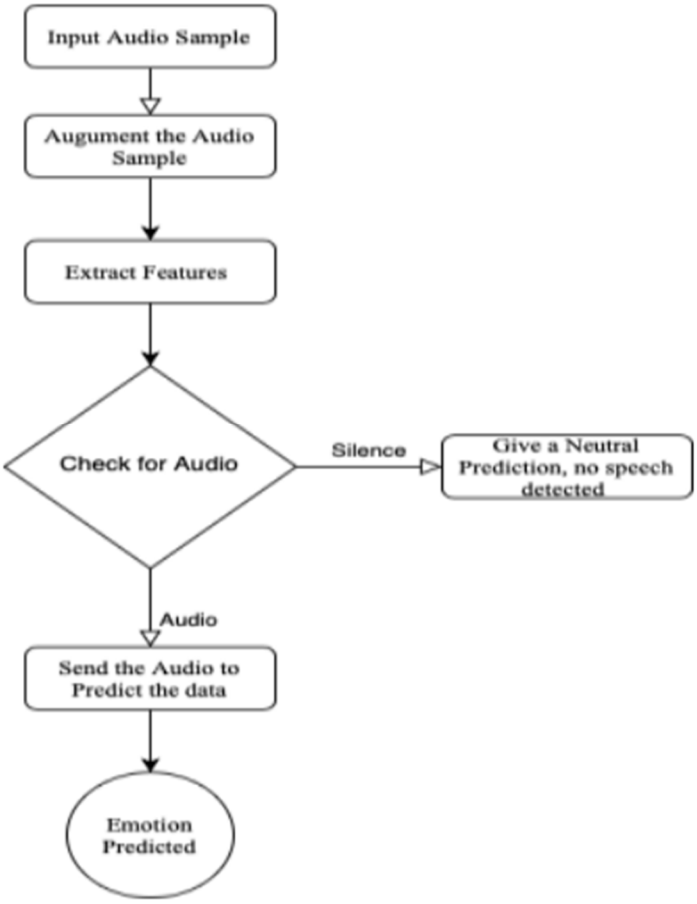


Fig. 1 Training Process Workflow

©IJRASET: All Rights are Reserved

Then we use another 25% of the remaining data set test to follow the feeling. Figure 3 shows the work flow of testing process.

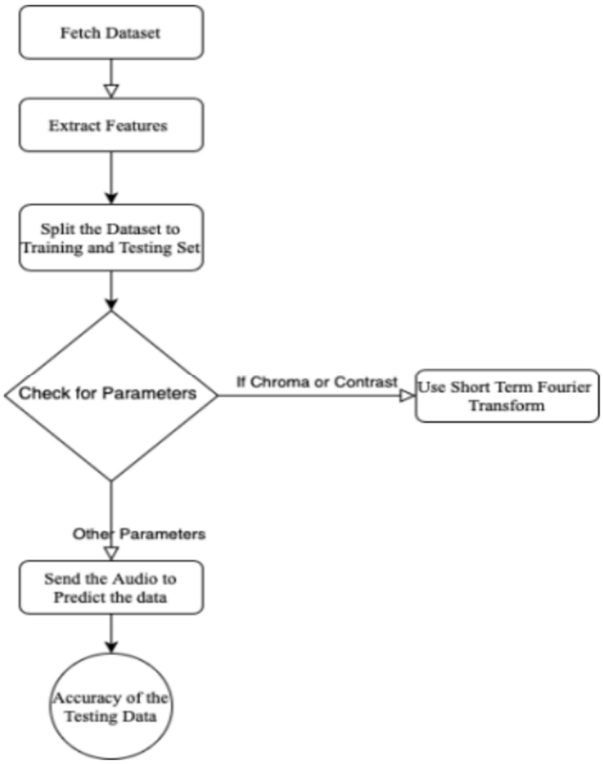


Fig. 3 Testing Process Workflow

The sound is recorded for 15 seconds, with the addition of 0.5 seconds gap at the beginning and end of the audio file to get the best catch the noise. Audio will differ from file volume, this will prevent the release of features, so that we avoid it and tend to measure the volume of it all sample. Audio length will be 32 Bit representation, as the number is represented in float format with ten signs in the manifestation of power, this is great value as it can represent large and large numbers, therefore increasing the width of the audio according to DB from -758 to 770 DB. MLP-Classifier takes the list of hyper parameters. Activation function used, is a divisive function that helps us find the slope of turns on any two points.

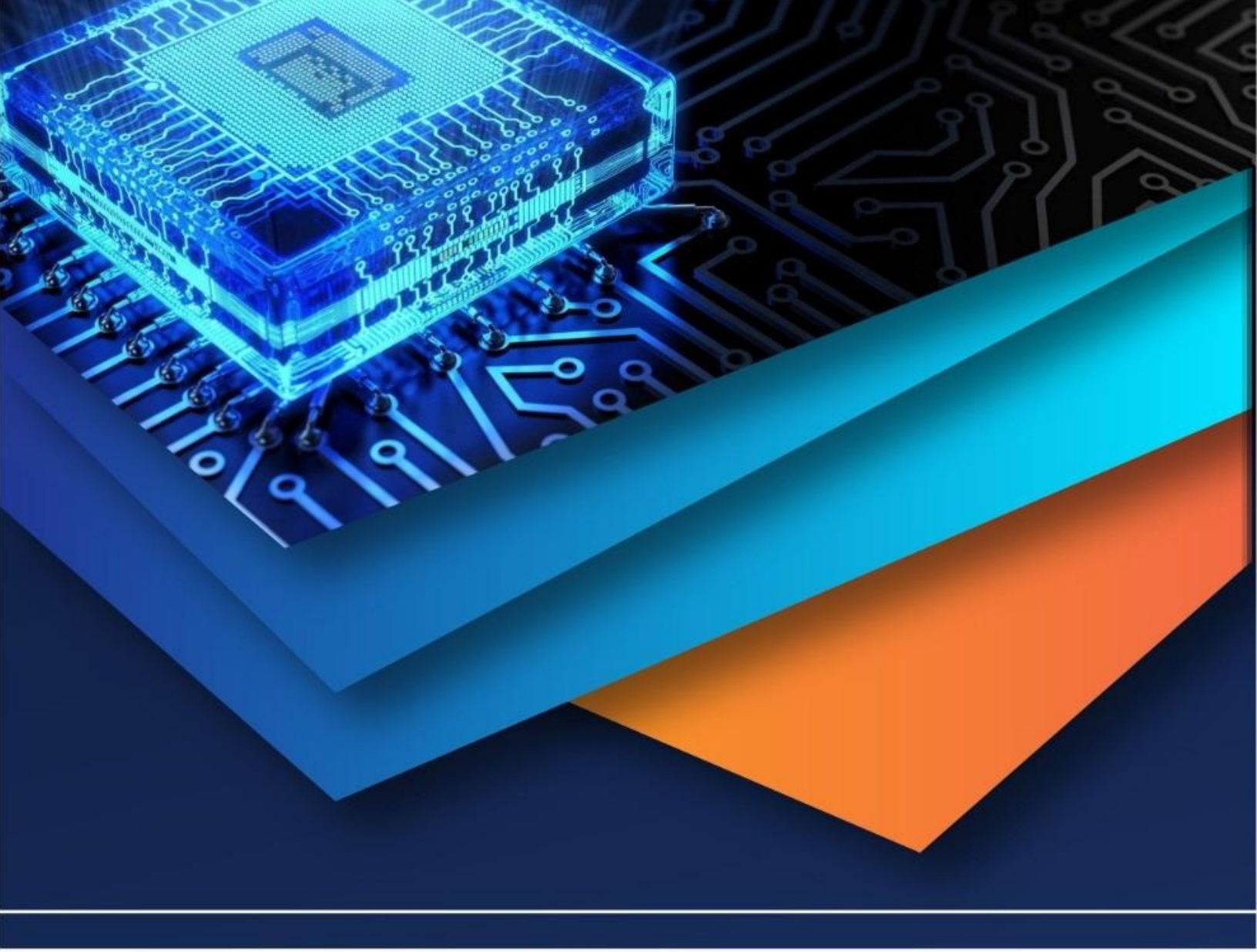
VI. CONCLUSIONS

This paper shows that MLPs have great potential for segregation speech signals. Even simplified models, a limited set of characters can be easily identified. We got high details compared to individual methods emotions. The performance of the module is highly dependent with pre-processing quality. Mel Frequency Cepstrum the most reliable coefficients. Every human emotion has it well-read, analused and accurate the results of this study show that speech recognition is possible, and that MLP's can be used any work related to the ability to speak and show the accuracy of each emotion present int the speech.

REFERENCES

[1] Navya Damodar, Vani H Y, Anusuya M A. Voice Emotion Recognition using Decision Tree. International Journal of Innovative Technology and Exploring Engineering (IJITEE), October 2019.
[2] Jianfeng Zhao, Xia Mao, Lijiang Chen. Learning Deep feature to Recognise Speech Emotion using Merged Deep CNN. IET Signal Process., 2018.
[3] H.K. Palo, Mihir Narayana Mohanty and Mahesh Chandra. Use of different features for Emotion Recognition using MLP network. Springer India 2015, Computational Vision and Robotics, Advances in Intelligent Systems and Computing.
[4] Ayush Kumar Shah ,Mansi Kattel,Araju Nepal. Chroma Feature Extraction using Fourier Transform. Chroma_ Feature extraction. January 2019.
[5] Sabur Ajibola Alim and Nahrul Khair Alang Rashid Some Commonly Used Speech Feature Extraction Algorithms.DOI: 10.5772/intechopen.80419.
[6] Davis, S.Mermelstein, P.(1980) Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences. In IEEE Transactions on Acoustics, Speech, and Signal Processing, Vol. 28 No. 4, pp. 357-366.
[7] X. Huang, A. Acero, and H. Hon. *Spoken Language Processing: A guide to theory, algorithm, and system development*. Prentice Hall, 2001.

©IJRASET: All Rights are Reserved



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Call : 08813907089 (24*7 Support on Whatsapp)