iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 10 **Issue:** V **Month of publication:** May 2022

DOI: https://doi.org/10.22214/ijraset.2022.42973

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: 7.538
Volume 10 Issue V May 2022- Available at www.ijraset.com

Speech Emotion Recognition Using Deep Learning

Vipul Kumar[1], Vibhor Varshney[2], Tushar Kushwaha[3], Dilkeshwar Pandey[4]
[1, 2, 3, 4]*CSE Department, KIET Group of Institutions, Ghaziabad, India*

Abstract: ***Emotional state identification based on analysis of vocalisations is a challenging subject in the field of Human-Computer Interaction (HCI). In the research that has been done on speech emotion recognition (SER). A wide range of research approaches has been used in order to extract feelings from a variety of inputs, including a number of well- known ways to speech analysis and categorization that are already known. Recent research has suggested the use of deep learning algorithms. as potential alternatives to the approaches that are traditionally used in SER. This article offers a summary of more in- depth topics learning methodologies, as well as current research employing it, are discussed to identify the feelings conveyed by verbal expressions. The analysis will consider the feelings that were recorded in the databases that were utilised were: the contributions to both speech and emotion that were removed the restrictions that were found, as well as the discoveries that were made discovered.***

Keywords: Speech emotions, Real-time Speech Classification, Transfer Learning, HCI Bandwidth Reduction, SER, LSTM

I. INTRODUCTION

The identification of speech emotions has developed from a specialised field application to an essential part of the human-computer interface participation (HCI)[1]. It is the goal of these systems to make natural Making the interaction between humans and machines more straightforward by interpreting information conveyed verbally via voice-to-voice contact rather than using conventional means of input and simplifying the process for users human listeners to whom a response may be given. Contact centre conversations, driver assistance systems built into automobiles, and the use of expressions of emotion gleaned from speech in medical .The term "spoken language" may be used to refer to apps as well conversation system [2]. Additional problems about HCI systems, However, we cannot ignore this, especially when it comes to these: The testing of systems moves from the lab to the actual world application. As a consequence of this, efforts are required to prevail overcoming these challenges and improving computer emotion recognition. [3]

There is a possibility that the quality of voice portals or contact centres will be evaluated using a detection method for rage. It makes it possible for service providers to tailor their deals to the feelings that their customers are now experiencing. Monitoring the stress levels of pilots may help lower the likelihood of an accident mishap with a plane in the field of civil aviation [4]. Many scholars have included the module for the detection of emotions into their merchandise in order to attract and keep a larger number of users interaction with video games. In order to improve the overall quality of cloud-based gaming information was used by Hossain et al. for their research distinct capacities for detecting and impacts on emotional states that sense emotions. The objective is to raise the level of involvement felt by players by customising the to their internal states of emotion. A psychiatric consultation using a chatbot .Within the domain of mental health, the provision of therapeutic services is recommended care[5].

Chatbot that engages in conversation and makes use of voice emotion Detection to facilitate conversation is yet another concept for an emotion recognition application. An active SER in real time The application should strive to do the most that is feasible the optimal combination of low processing load and high throughput times, as well as a high level of accuracy.

The most important parts of the speech are the feelings. The techniques of speech recognition (SER) include feature extraction and feature classification. Excitation based on the source characteristics, as well as prosodic characteristics, have all been created by experts in the study of speech by academics in the area processing. The second step is to combine the ingredients element that separates data using both direct and indirect separators.

Bayesian Networks, often known as BN, or the Maximum Likelihood Support Vector Machine (SVM) and the Likelihood Principle (MLP) Machine are two of the linear technologies that are used most often criteria for the identification of feelings and emotions. It was the voice signal is often regarded to be moving in a non- stationary fashion. As a direct consequence of this, non-linear classifiers are thought to achieve their goals for SER.

©IJRASET: All Rights are Reserved

The Gaussian and the spectral energy ratio (SER) are two non-linear classifiers. The Mixture Model, abbreviated as GMM, and the Hidden Markov Model Model (HMM) (HMM). The classification of data often makes use of these discovered in the most fundamental aspects. power-based characteristics such is Linear Predictor Coefficients (LPC), Mel Energyspectrum Dynamic Coefficients (MEDC), and Mel Energyspectrum Coefficients Frequency Cepstral Coefficients, often known as MFCC, in addition to Cepstral coefficients derived from the Perceptual Linear Prediction model (PLP) are often employed as an efficient tool for identifying feelings from audio. Other descriptors that may be used for feelings

K-Nearest Neighbour is included in the identification (KNN),Decision Making and Principal Component Analysis (PCA) trees.

II. OBJECTIVE

The basic goal of SER is to advance human health and well- being interaction with a machine. It is also employed in the act of lying sensors that can follow the movements of a person condition of the mind [6]. Words and emotions detection has recently seen an uptick in its prevalence in the both the medical and forensic scientific communities. Pitch and timbre The characteristics of prosody are used to identify seven distinct feelings across the course of this activity.

III. PROPOSED WORK

A number of investigations have been carried out to sift views toward audio data. Recognizing one's feelings Taking features from speech involves a number of steps based on a collection of highly emotive texts, the phrase used, and then organise the feelings that it conveys based on the findings that were collected. [7] The exactness with which The manner in which the traits were developed has a significant influence about the ability to organise one's feelings.

In their proposal, Noroozi and colleagues proposed a versatile a method of identifying feelings that uses both visual and auditory cues processing of sounds heard using the ears. According to his findings, PCA is used in order to diminish the impact of previously 88 characteristics were used to determine which ones were omitted (MFCC), During the element extraction process (using FBEs) (PCA).

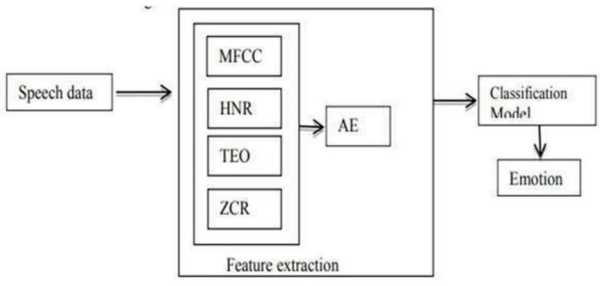
The GMM classifier was used by Bandela in order to identify five using the Berlin Emotional Speech to express one's database by including a sonic characteristic referred as as the MFCC that has a rhythmic element known as the Teager Energy Operator. Within the framework that they proposed, Spectral features were used in Zamil et alstudy the 13 MFCC that are extracted from the audio data are as follows:

in order to identify all eight feelings by employing the LSTM method with a degree of precision that is equivalent to ninety percent. [8]

In addition to this, the precision cannot be maintained in such a manner over ninety percent, which might have an effect on the capacity to recognise the feelings conveyed by a speaker's words. Multiple According to the opinions of doctors, the most prevalent auditory factors to consider in selecting a Teager Energy Operator (TEO), the zero crossings Rate (ZCR), the MFCC, and the filter bank's energetic properties and attributes .All feelings may be categorised as energies (FBE). We make available a method for detecting feelings that is based on Support Vector Machines that have the features number 39 In this study, MFCC, HNR, ZCR, and TEO are referred to as [9]. We are going to examine the performance of two different systems.

Use an auto-encoder to pull out the pertinent information. criteria based on previously obtained parameter information then classifying them based on the support vector machines.

IV. PICTORIAL REPRESENTATION



A. Feature Extraction

We utilised 39 MFCC (12 MFCC + energy, 12 delta MFCC +energy) within the scope of this research, in addition to the Zero Crossing Rate, Both the Teager Energy Operator and the Harmonic to Noise Ratio are included.

```
[18]: extract_mfcc(df['speech'][0])

[18]: array([-286.02704  ,   86.23414  ,   -2.635008 ,   22.56944  ,
             -15.209238 ,   11.531056 ,   11.94983  ,   -2.5640693 ,
               0.63499874,   11.539477 ,  -17.854872 ,   -7.673544  ,
               6.1565223 ,   -3.8004475 ,   -9.552902 ,    3.9220483 ,
             -13.588674 ,   14.449348 ,   19.316969 ,   23.08124   ,
              32.217903 ,   16.650953 ,   -4.138859 ,    1.20185   ,
             -11.535382 ,    6.9179306 ,   -2.8407633 ,   -7.467058  ,
             -11.169154 ,   -2.1890635 ,   -5.489987 ,    4.463149  ,
             -11.367347 ,   -8.866391  ,   -3.8222973 ,    5.0004168 ,
              -1.7143741 ,    2.6642315 ,   11.361882  ,   11.313575  ],
            dtype=float32)
```

```
[21]: X_mfcc

[21]: 0       [-286.02704, 86.23414, -2.635008, 22.56944, -1...
      1       [-348.74265, 35.7786, -4.4302225, 15.252251, 5...
      2       [-340.4982, 54.36257, -14.845929, 21.453777, 8...
      3       [-307.126, 21.987495, -5.146962, 7.2266817, -8...
      4       [-345.27826, 47.107338, -24.942877, 20.17883, ...
                                    ...
      5595    [-374.88943, 61.58835, -0.7048707, 9.159214, -...
      5596    [-314.5199, 40.65166, -6.461393, -3.05994, -51...
      5597    [-358.06696, 78.64016, -15.999416, 2.9611573, ...
      5598    [-353.46466, 102.15106, -14.645692, -11.564197...
      5599    [-389.82825, 54.579174, 0.8075471, -0.8762022,...
      Name: speech, Length: 5600, dtype: object
```

B. Feature Dimension Reduction

The practise of decreasing the proportions of certain features while preserving as much of the essential information is possible as practically possible is referred to as the feature dimension decrease. There are two methods that may be used to the feature size is comprised of the feature selection as well as the feature removal. The auto- encoder (AE) is what determines which characteristics are shown inquiry into the matter.

This strategy is quite similar to the conventional one artificial neural network has several layers, in the sense that itis a non-recurrent, feed-forward neural circuit network .Learning a new skill is the point of incorporating AE into the task less revealing data display.

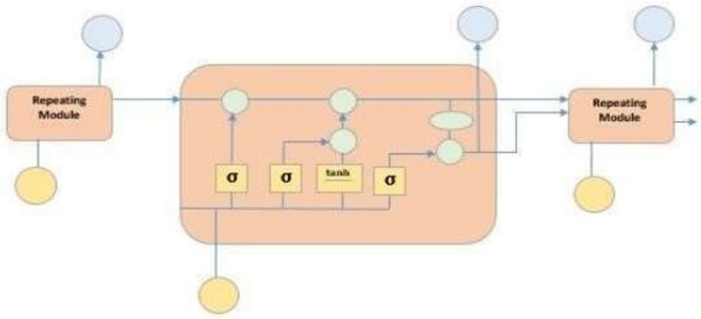
This system possesses a three layers: an input/output layer, an output/processing layer, and one or more levels additional hidden layers .[15]The present investigation makes use of the and kinds of auto-encoders.AE fundamentals in addition to the stacking auto-encoder.[9]The total number of concealed layers varies depending on the two different kinds; the standard AE only has one, whereas the Layered autoencoders consist of two or more than two encoders.

C. Classification Model

SVMs were first developed for the purpose of discrimination between two different groups. Expanding a subject in a variety of ways split-split splitter into multi-stage split. There have been several application developments. Multi-class SVMs find applications in many different fields of business and have shown to be beneficial in the classification of a wide range of sources of information .[18]In order to solve the issue, the SVMs will first investigate a decomposition that consists of quite a few binary components classifiers. SVMs, or support vector machines, are a kind of machine learning supervised machine learning and has applications in data classification and forecasting are also included. It makes an effort to Acquiring a hyperplane that can be used to categorise materials is the first step divide the data into really huge genes, which will serve you very well identifies the many pieces of training data that are presented in the feature area by the K function of the kernel, which is the kernel that is utilised the most often A variety of functions, including linear, polynomial, and RBF, are used in order toto categorise new values based on the training dataset and conduct analysis on them. As a consequence of this, the only choice available to utilise this partition is to discover the relevant kernel functions and adjust the settings to fit your needs get the maximum level of detection that is achievable. [10]Then, if we switch the kind of SVM kernel, we will do the following :based on actual data, make adjustments to the SVM classifier's many parameters.in order to determine the variables r, d, and c, which are chosen by the user values, in order to determine which kernel parameters are optimal for our system search.

D. LSTM Model

It is an advanced kind of recurrent neural network that has the ability to learn long-term data trends. Due to the fact that the model is repeated module is composed of four tiers that communicate with one another. This is not impossible if we work together on it .[19]The graphic that can be seen above depicts four layers of neural networks in yellow input is represented in yellow boxes, point-wise operators in green circles, and output in blue circles, and cell state in blue circles. This is the LSTM module consisting of three gates and a cell structure, this construction gives you the ability to read just some of the information from each unit, leave some of it unread, or store it.[17] The state of the cell in LSTM determines whether or not data may travel through it.by preventing most modifications to the units and permitting just a handful of them interactions. Each individual unit is equipped with an I/O, an O/P, and a forget gate for data addition and subtraction based on the current cell state. The forget gate makes its determinations with the help of a sigmoid function whether or whether information from the prior state of the cell should be discarded. in order to control the flow of information to the current cellstate, the input gate will carry out a point-wise multiplication of the points operation of 'sigmoid' and 'tanh.' Finally, the output gate choose one of these two which information should be sent on to the next concealed state.[20]

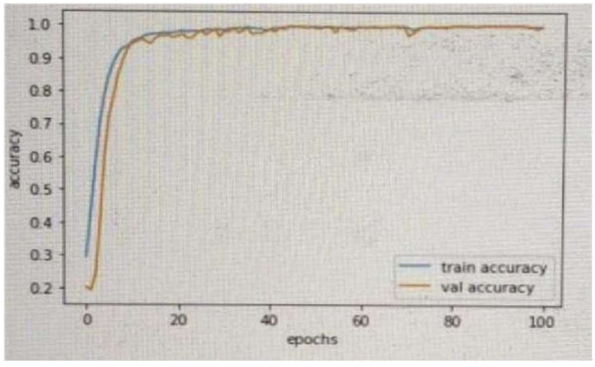


Layer (type)	Output Shape	Param #
lstm (LSTM)	(None, 123)	61500
dense (Dense)	(None, 64)	7936
dropout (Dropout)	(None, 64)	0
dense_1 (Dense)	(None, 32)	2080
dropout_1 (Dropout)	(None, 32)	0
dense_2 (Dense)	(None, 7)	231

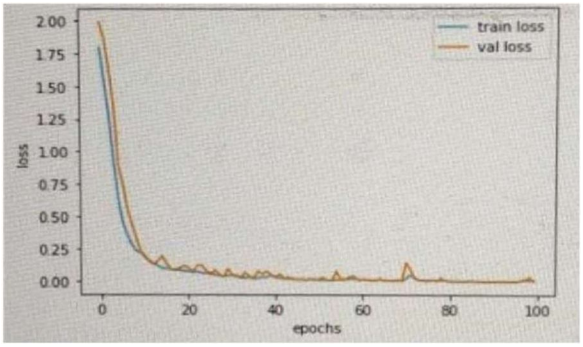
Total params: 71,747
Trainable params: 71,747
Non-trainable params: 0

V. RESULT

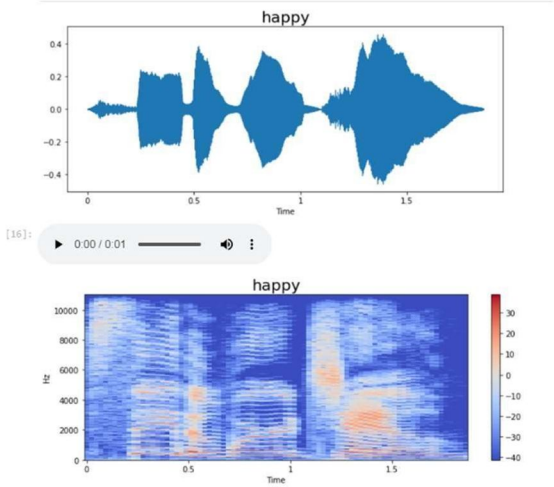
1) It is the accuracy graph which is showing very good accuracy.



2) It is the loss graph which is showing very less loss.



The below image is the result for one emotion happy with waveshow plot.



VI. CONCLUSION

The purpose of this study is to provide a general perspective on recent developments in the ability to recognise emotions from speech systems. Finding new and improved solutions is the mission of the SER ways to extract emotion from audio. Utilization of designs of deep convolutional learning that are capable of learning from Presentations of voice spectrograms are becoming increasingly common. They are widely acknowledged as a reliable foundation for SER systems, in addition to recurrent network structures. More intricate the designs of SERs have evolved throughout the course of time, with a concentrate on obtaining emotionally meaningful information from the worldwide circumstances. According to the results of our research, the focus mechanism is capable of assisting SER Improved system performance may not always manifest itself in observable ways.

REFERENCES

[1] Yenigalla, P.; Kumar, A.; Tripathi, S.; Singh, C.; Kar, S.; Vepa, J. Speech Emotion Recognition Using Spectrogram & Phoneme Embedding. In Proceedings of the INTERSPEECH, Hyderabad, India, 2–6 September 2018.

[2] Koolagudi, S.G.; Murthy, Y.V.S.; Bhaskar, S.P. Choice of a classifier, based on properties of a dataset: Case study-speech emotion recognition. Int. J. Speech Technol. 2018, 21, 167–183.

[3] Zhang, S.; Zhang, S.; Huang, T.; Gao, W. Speech Emotion Recognition Using Deep Convolutional Neural Network and Discriminant Temporal Pyramid Matching. IEEE Trans. Multimed. 2018, 20, 1576– 1590.

[4] Xi, Y.; Li, P.; Song, Y.; Jiang, Y.; Dai, L. Speaker to Emotion: Domain Adaptation for Speech Emotion Recognition with Residual Adapters. In Proceedings of the 2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), Lanzhou, China, 18–21 November 2019; pp. 513–518

[5] B. W. Schuller, "Speech emotion recognition: Two decades in a nutshell, benchmarks, and ongoing trends," Communications of the ACM, vol. 61, no. 5, pp. 90–99, 2018.

[6] M. S. Hossain and G. Muhammad, "Emotion recognition using deep learning approach from audio– visual emotional big data," Information Fusion, vol. 49, pp. 69–78, 2019.

[7] A. B. Nassif, I. Shahin, I. Attili, M. Azzeh, and K. Shaalan, "Speech recognition using deep neural networks: A systematic review," IEEE Access, vol. 7, pp. 19

143–19 165, 2019.

[8] Fayek, H.M.; Lech, M.; Cavedon, L. Evaluating deep learning architectures for Speech Emotion Recognition. Neural Netw. 2017, 92, 60–68.

[9] Tzinis, E.; Potamianos, A. Segment-based speech emotion recognition using recurrent neural networks. In Proceedings of the 2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII), San Antonio, TX, USA, 23–26 October 2017; pp. 190–195.

[10] J. Han, Z. Zhang, F. Ringeval, and B. Schuller, ''Reconstruction-errorbased learning for continuous emotion recognition in speech,'' in Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP), Mar. 2017, pp. 2367–2371.

[11] G. Wen, H. Li, J. Huang, D. Li, and E. Xun, ''Random deep belief networks for recognizing emotions from speech signals,'' Comput. Intell. Neurosci., vol. 2017, Mar. 2017, Art. No. 1945630.

[12] L. Zhu, L. Chen, D. Zhao, J. Zhou, and W. Zhang, ''Emotion recognition from Chinese speech for smart affective services using a combination of SVM and DBN,'' Sensors, vol. 17, no. 7, p. 1694, 2017.

[13] P. Tzirakis, G. Trigeorgis, M. A. Nicolaou, B. Schuller, and S. Zafeiriou, ''End-to-end multimodal emotion recognition using deep neural networks,'' IEEE J. Sel. Topics Signal Process., vol. 11, no. 8, pp. 1301–1309, Dec. 2017.

[14] Khalil, R.A.; Jones, E.; Babar, M.I.; Jan, T.; Zafar, M.H.; Alhussain, T. Speech Emotion Recognition Using Deep Learning Techniques: A Review. IEEE Access 2019, 7, 117327–117345. [CrossRef]

[15] Akçay, M.B.; O ˘guz, K. Speech emotion recognition: Emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers. Speech Commun. 2020, 116, 56–76.

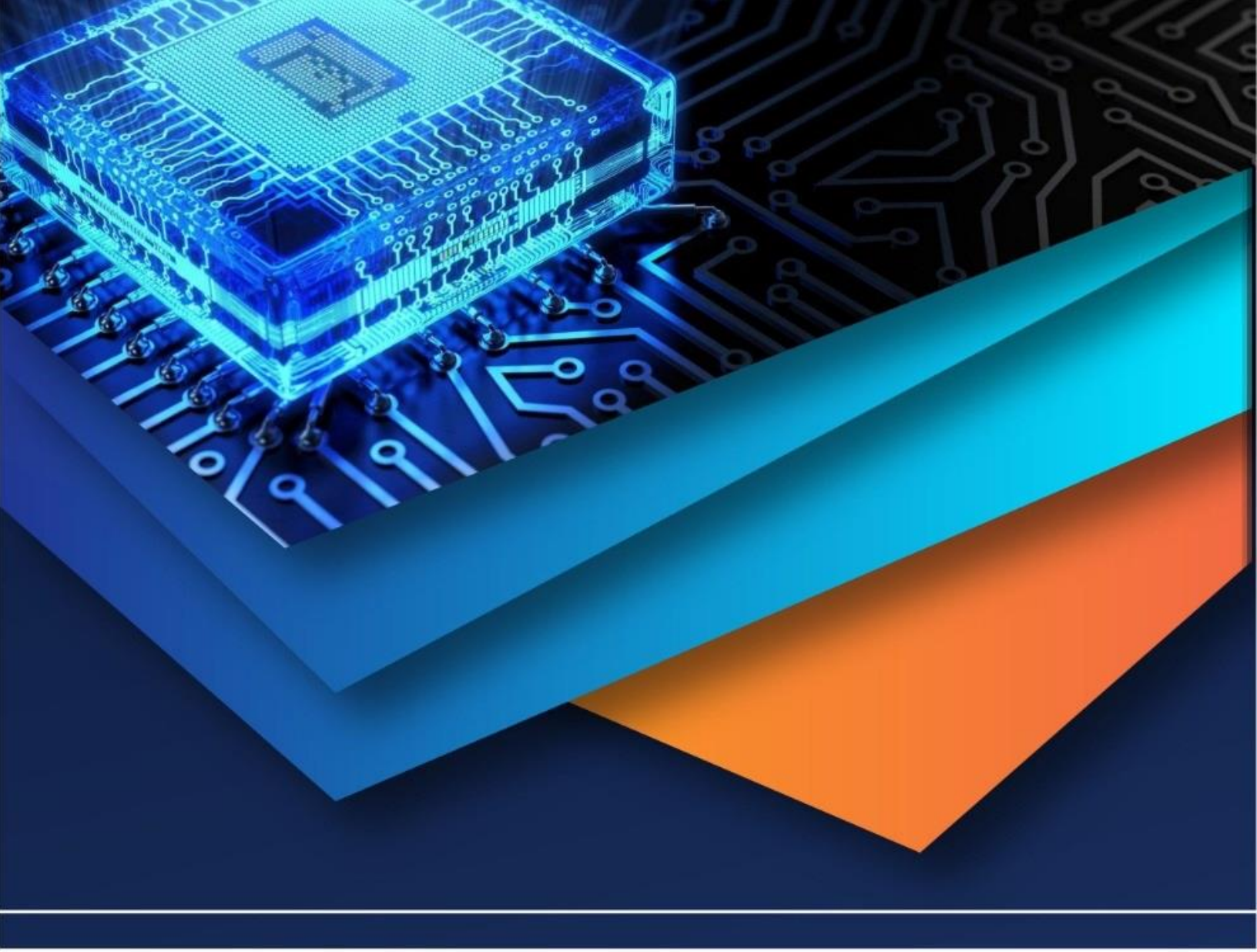
[16] Avots, E., Sapiński, T., Bachmann, M. Et al. "Audiovisual emotion recognition in wild". Machine Vision and Applications (2018).

[17] L. Kerkeni, Y. Serrestou, M. Mbarki, K. Raoof, M. Ali Mahjoub and C. Cleder. "Automatic Speech Emotion Recognition Using Machine Learning". Social Media and Machine Learning. (2019).

[18] Zamil, Adib Ashfaq A., et al. "Emotion Detection from Speech Signals using Voting Mechanism on Classified Frames." 2019 International Conference on Robotics, Electrical andSignal Processing Techniques (ICREST). IEEE, 2019.

[19] S.B. Reddy, T. Kishore Kumar , "Emotion Recognition of Stressed Speech using Teager Energy and Linear Prediction Features," in IEEE 18th International Conference on Advanced Learning Technologies, 2018.

[20] Schuller, B.W. Speech emotion recognition: Two decades in a nutshell, benchmarks, and ongoing trends. Commun. ACM 2018, 61, 90–99



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)